Aprendizaje automático y modelización de tópicos: un estudio de caso sobre la agenda mediática en contexto de las elecciones Argentina 2015

Tesina para obtener el título de Licenciado en Sociología.

Carrera de Sociología.

Escuela Interdisciplinaria de Altos Estudios Sociales. UNSAM.

Estudiante: Tomás Maguire.

Director: Dr. Germán Rosati.

Fecha: Abril 2021.

Aprendizaje automático y modelización de tópicos: un estudio de caso sobre la agenda mediática en contexto de las elecciones Argentina 2015

Autor

Tomás Maguire

Director

Dr. Germán Rosati

Evaluador

Resumen

Este trabajo se propone estudiar la conformación de la agenda mediática en contexto de las elecciones presidenciales llevadas a cabo en Argentina en el año 2015. Para ello, se propone la implementación de técnicas computacionales en función de recolectar, procesar y analizar las noticias publicadas por los medios *Clarín, La Nación, Página 12, MinutoUno, Télam, Perfil e Infobae* entre enero de 2015 y diciembre de 2016 inclusive. Se propone indagar sobre cuál es el efecto de la coyuntura electoral en la agenda mediática de los medios mencionados, así también describir la cobertura que se le da a los diferentes tópicos en el recorte espacio – temporal planteado.

Se definió utilizar una técnica que proviene del campo del aprendizaje automático: la modelización de tópicos. Se optó por implementar el modelo *Latent Dirichlet Allocation (LDA)* con el objeto de detectar los tópicos latentes que existen sobre el corpus de texto generado a partir de las diferentes definiciones metodológicas sobre minería de texto respecto a la recolección, preprocesamiento y disposición de los datos de la investigación, en este caso noticias periodísticas.

En el primer capítulo se expone el marco teórico, el estado del arte y la metodología. El marco teórico toma elementos principalmente de la teoría de *agenda setting* y otros estudios de medios que estudian cuál es la influencia de los tópicos que abordan los medios de comunicación sobre el discurso público o las agendas de los ciudadanos. Se exponen también trabajos realizados sobre estudios de medios de comunicación que tuvieron por objeto detectar temas o tópicos en diferentes recortes de tiempo y con diferentes metodologías. También se realiza un breve recorrido histórico sobre la subdisciplina de las ciencias sociales computacionales y el conjunto de técnicas computacionales denominadas *machine learning* o aprendizaje automático.

El segundo capítulo explica técnicas de procesamiento de lenguaje natural para luego detallar en qué consiste tanto la modelización de tópicos como la intuición matemática detrás del modelo *LDA* y la distribución Dirichlet.

Por último, el tercer capítulo hace referencia a la implementación del modelado de tópicos de forma completa, desde el proceso de ingeniería de datos, la minería de texto y la implementación de la técnica escogida de aprendizaje automático. Como principal hallazgo es posible mencionar el haber constatado un cambio en los tópicos más relevantes coincidente con el calendario electoral de 2015. También fue posible confeccionar visualizaciones para describir de manera detallada la cobertura de los tópicos predominantes durante el recorte espacio temporal que se ha especificado.

Agradecimientos

En primer lugar, agradecer a mi director Germán Rosati. Cualquier línea que pueda redactar siento que es insuficiente para expresar la gratitud por su acompañamiento, consejos y ayuda en este recorrido. La disciplina de las ciencias sociales computacionales me resultó (y resulta) un fuerte desafío, tuve la fortuna de encontrarme con una persona que hizo de ese recorrido algo sumamente ameno. Mi mayor admiración y afecto hacia él.

Agradecer también a Analía y Florencia por sus lecturas cuidadosas y excelentes sugerencias, muchas gracias por haber destinado su tiempo a este trabajo.

A mis compañeras y compañeros de cursada, docentes, trabajadoras y trabajadores de la universidad. Mi mayor gratitud hacia cada una de esas personas, por su calidez, por su trato y predisposición.

A la universidad pública, laica y gratuita que permitió que un laburante más pueda tener una educación de una institución de excelencia.

A mis amigas y amigos por su compañía, consejos y afecto.

Por último, y principalmente, a mi viejita. Por tu ejemplo y tu guía. Por tu solidaridad y tu entereza.

"Nací en un barrio donde el lujo fue un albur

Por eso tengo el corazón mirando al sur

Mi vieja fue una abeja en la colmena

Las manos limpias, el alma buena.

Y en esa infancia, la templanza me forjó

Después la vida mil caminos me tendió

Y supe del magnate y del tahúr

***Por eso, por eso tengo el corazón mirando al sur**"*

El corazón mirando al sur – Eladia Blázquez

INDICE GENERAL

Índice de fotografías

Índice de cuadros

Índice de gráficos

.....

El código e indicaciones de implementación del presente trabajo se encuentra disponible en **https://tomasebm.github.io/topicmodeling/**

.....

Introducción

Esta tesis tiene por objetivo estudiar la conformación de la agenda mediática en el periodo electoral de 2015 en Argentina. Para ello se propone implementar una técnica de modelización de tópicos proveniente del campo del aprendizaje automático[1]. En este marco, se procura responder los siguientes interrogantes: ¿cuáles son los tópicos relevantes que tratan los medios de comunicación seleccionados en la coyuntura electoral del año 2015 en Argentina? ¿Cuál es su evolución temporal? ¿Cuál es la utilidad de las técnicas (particularmente, las vinculadas al modelado de tópicos) para el estudio de estas cuestiones?

Como explica Rosati (2020), el trabajo empírico de las ciencias sociales se caracteriza por una gran diversidad de fuentes de información, desde datos altamente estructurados hasta datos de menor grado de estructuración, son estos últimos los que ocupan un lugar central. Datos textuales como documentos, revistas, noticias, entrevistas, son insumos frecuentes en la disciplina de las ciencias sociales. Las herramientas y técnicas metodológicas de las que se suele hacer uso se suelen encontrar más vinculadas al análisis literario o del discurso y a métodos que buscan la comprensión profunda de los corpus. Las posturas interpretativistas tienden a poner el eje en la capacidad subjetiva del investigador para realizar la interpretación o el análisis. "Esto hace que parte de las investigaciones puedan presentar una relativa falta de sistematicidad metodológica. Al mismo tiempo, el peligro de la imposibilidad de replicación de sus resultados suele estar latente" (Rosati, 2020:2).

La combinación de técnicas como la minería de textos y el procesamiento de lenguaje natural pueden ser de utilidad atendiendo diferentes dificultades propias de la adopción de procedimientos de control metodológico tradicionales, ya sean analógicos o mediatizados por software. Los procesos tradicionales continúan teniendo un carácter fuertemente manual y continúan residiendo en el investigador. Bases manuales, interfaces gráficas, inexistencia de scripts o sintaxis que codifiquen las operaciones muestran que los procesos aún poseen un grado medio o bajo de estandarización. La posibilidad de replicar el procesamiento como el análisis se encuentra atada a la decisión del investigador de documentar de manera minuciosa cada una de las decisiones tomadas. Se adicionan problemas como la escalabilidad: la

[1] A grandes rasgos podemos definir el aprendizaje automático como sistemas computacionales que aprenden y se adaptan sin seguir instrucciones explícitas utilizando algoritmos y modelos estadísticos para analizar e inferir patrones en los datos objeto del análisis. Se abordará el concepto en detalle en el capítulo II.

necesidad de transcripción y codificación manual de corpus de texto (aún con la asistencia de software) limita fuertemente el tamaño del corpus que es posible analizar.

La combinación de técnicas propuestas en este trabajo puede ser de utilidad a las ciencias sociales atendiendo los problemas anteriormente expuestos. Su uso debería permitir:

a) Sistematizar (eventualmente logrando un grado de automatización) diversas etapas del proceso de investigación, desde la recolección de datos, a la construcción del corpus, su preprocesamiento y análisis.

b) Aplicar métodos cuantitativos, específicamente técnicas de procesamiento de lenguaje natural, habilitando una amplia diversidad de tareas (clasificación de textos, detección de temas y tópicos, detección de estructuras semánticas, etc.).

c) Escalabilidad: en lugar de leer cada uno de los textos de un corpus, la técnica de minería de texto permite analizar de forma automática corpus de texto de gran escala.

El presente trabajo intenta ser una aproximación metodológica al análisis computacional de textos y presenta algunas de sus potencialidades posibles de ser implementadas en las ciencias sociales.

El recorte temporal se extiende desde enero de 2015 a diciembre de 2016 con la intención de incluir el inicio del año electoral y un año posterior para intentar visualizar si el ordenamiento de tópicos relevantes exhibe alguna modificación luego del acto electoral. Los medios incluidos en el análisis son: *Clarín*, *La Nación*, *Página/12*, *Infobae*, *Perfil*, *MinutoUno* y *Telam*[2]. El efecto de la agenda mediática sobre la discusión pública es un tema recurrente en los estudios de medios y comunicación, en este trabajo se optó por tomar algunos elementos fundamentales de la teoría de *agenda setting*. Exponiéndolo de forma sintética dicha teoría plantea, según explica uno de sus iniciadores Max McCombs que "elementos prominentes en los medios de comunicación con frecuencia adquieren prominencia entre el público" (citado en Aruguete, 2015). Asimismo, si bien se adopta algunos elementos de tal teoría para fundamentar los motivos del estudio de los grandes medios de comunicación y su producción periodística, el

[2] No existe un consenso unívoco respecto a la metodología de selección de medios de comunicación. Optamos por guiar la elección tomando en consideración el ranking de sitios Alexa.com y la consultora ComScore (Becerra, 2019).

trabajo se centra en la aplicación de un modelo estadístico para el estudio de los principales temas que proponen los medios periodísticos mencionados en el recorte espacio temporal propuesto. En otras palabras, el foco está puesto en detectar esos elementos prominentes en los medios de comunicación masiva que menciona McCombs.

Para responder los interrogantes planteados, se realizará una modelización de tópicos utilizando el modelo *LDA (Latent Dirichlet Allocation)* obteniendo del mismo una matriz de datos que agrupará los tópicos recurrentes (o "latentes" en términos del modelo) del corpus de texto que se le suministre al mismo. El proceso consta de una consulta a la base de datos GDELT[3] con las especificaciones mencionadas respecto a medios requeridos y fechas propuestas, luego el resultado de la consulta es procesado por una aplicación desarrollada en Python que *scrapea*[4] y almacena el corpus, título y copete en un set de datos maestro. Al mismo se le realiza una serie de procesos de limpieza, con la intención de generar el set de datos óptimo para ser volcado al modelo generativo *LDA*. Luego, habiendo obtenido la matriz de datos, se analizará, graficará y se confirmará o rechazará la hipótesis planteada por la investigación. Se intenta demostrar que el contexto pre y poselectoral inciden en la conformación y el tipo de cobertura que reciben los diferentes tópicos relevantes en los medios de comunicación estudiados.

Esta tesis se organiza en tres capítulos. En el primer capítulo se expondrá el estado del arte y el marco teórico. Se realizará una revisión de la formación histórica de los conceptos claves de las ciencias sociales computacionales, para luego explorar la metodología de aprendizaje automático. En el segundo capítulo se abordará conceptos claves del proceso de minería de texto y procesamiento de lenguaje natural. Por último, en el tercer capítulo se procederá a implementar el modelo propuesto sobre los datos recabados y presentar los principales resultados.

Como principales conclusiones que se desprenden del presente trabajo se menciona a) la efectiva verificación de un cambio de los tópicos más relevantes en una coyuntura electoral, detectados mediante la implementación del modelo LDA sobre el corpus de texto objeto de

[3] GDELT (*Global Database of Events, Language and Tone*) se presenta como un proyecto respaldado por Google Jigsaw que monitorea la emisión de noticias web de casi todos los rincones del mundo en más de 100 idiomas, al mismo tiempo que identifica personas, localidades, organizaciones, temas, fuentes, emociones, conteos, citas, imágenes y eventos, creando una plataforma abierta y gratuita para computar. Para más detalles: https://www.gdeltproject.org/

[4] El término hace referencia a la automatización del proceso de insertar en una base de datos la información requerida de un sitio web que contiene una noticia.

análisis, b) la confección de visualizaciones que dan cuenta de la evolución de la composición de la media de los tópicos de noticias, logrando observar particularidades en la evolución de tópicos al desagregar el análisis por medio de comunicación.

Capítulo I: Marco teórico, estado del arte y métodos computacionales

Los medios de comunicación no son agencias de información neutrales. Por el contrario, son "actores políticos" con intereses y metas que procuran encontrar resonancia de sus opiniones tanto en las audiencias como en la política (Eilders 1997, 2000). De esta manera, los medios de comunicación establecen los temas relevantes de discusión en la esfera pública. Este proceso de definición temática se ha visto complejizado en las últimas décadas a la luz del nacimiento y la posterior expansión de los llamados "medios 2.0" o medios digitales.

En principio, la cantidad de información existente previo al auge de los medios digitales ya presentaba un desafío para los científicos que pretendían abordar tal objeto de estudio, dada la gran cantidad de información generada constantemente por múltiples medios de comunicación. En la actualidad, esta información no solo aumentó en su tamaño de forma considerable, sino que también se ha visto ampliada en su heterogeneidad. Más medios, más voces, y más espacios de comunicación virtuales, presentan un desafío desde el punto de vista del análisis, procesamiento y visualización de los datos que se generan día a día por millones de usuarios y organizaciones.

Como se mencionó anteriormente, mediante la selección y el énfasis en ciertos temas, los medios expresan su "particular posición política" desde un perfil ideológico que los distingue (Eilders, 2000: 181). En este sentido, la noción clave de la teoría de agenda setting radica en la transferencia de la prominencia de la agenda mediática hacia la agenda del público y esta noción, explica Max McCombs, es utilizada en escenarios cada vez más diversos a medida que el panorama de los medios se expande para incluir a toda una gama de redes sociales (Como se cita en Aruguete, 2015).

I.1 Marco teórico

McCombs señala como padre intelectual de la teoría de *agenda setting* a Walter Lippman, quien en cuyo clásico "La opinión pública" (1922) ya indagaba e intentaba trazar cuál era la influencia de los medios de comunicación sobre "nuestras imágenes mentales". McCombs inscribe entonces sus trabajos iniciales (1968) y posteriores, en una continuación del esfuerzo inicial de Lippman por dar cuenta de tal relación (como se cita en Aruguete, 2015). Los primeros estudios de Maxwell McCombs y Don Shaw en Chapell Hill plantea tres resultados principales, que la podemos resumir en: a) el alto grado de correspondencia entre el patrón de

cobertura de noticias sobre asuntos públicos y la percepción de la gente sobre los temas más importantes del día; b) nominar la relación entre agenda mediática y agenda pública como agenda setting o configuración de la agenda, el cual comunica la esencia de la teoría en una manera fácil de comprender; c) la forma de medir la relación precisa entre la agenda mediática y la agenda pública es la correlación de orden de prioridades (Aruguete, 2015: 15). Zunino y Grili Fox (2019) explican que las agendas mediáticas son el resultado de intensos procesos productivos de selección, omisión y jerarquización que se dan en las redacciones, a partir de los cuales los medios de comunicación estructuran una propuesta temática otorgando relevancia a algunos asuntos en detrimento de otros. El concepto de tema es definido como una serie de acontecimientos relacionados con el tratamiento periodístico que se agrupan en una categoría más amplia. Estos acontecimientos, directamente observables en la superficie del discurso, constituyen tópicos, es decir, etiquetas que resumen el dominio de las experiencias sociales incluidas en el relato (Pan & Kosicki, 1993).

Calvo (2015) advierte sobre los sesgos y limitaciones que trae a cuestas la realización de análisis sobre redes sociales y, tomando como caso Twitter, hace hincapié en fenómenos como el efecto cámara de eco, dado que los algoritmos que ordenan la información dispuesta a un usuario tienden a mostrarle información que generalmente es consistente con sus prejuicios. En efecto, no hay una respuesta unívoca que indique si los nuevos medios les disputan la agenda a los medios tradicionales o si repiten su temario, menos aún si tienen la capacidad de establecer la agenda pública (Aruguete, 2015).

La modelización de tópicos, desarrollada en el campo de las ciencias de la computación, el aprendizaje automático y el procesamiento de lenguaje natural, identifica tópicos y captura "la relacionalidad del significado" (DiMaggio, 2013), se utilizará en este trabajo el modelo *Latent Dirichlet Allocation (LDA)*. La intuición detrás de este modelo es que cada documento que pertenece a un corpus de texto puede exhibir varios tópicos simultáneamente, esto es, hablar de varios temas al mismo tiempo. Por ejemplo, una noticia que hace referencia a las elecciones PASO de 2019 puede hacer referencia a varios temas en el mismo texto: inseguridad, corrupción, intención de voto, incluso deportes o farándula. La redacción de una noticia usualmente aborda diferentes aristas de la realidad. La idea detrás de LDA es operacionalizar esta intuición a través de un modelo generativo, es decir, asumiendo la existencia de un "proceso generador de textos": un proceso aleatorio imaginario por el cual un documento es producido. El objetivo entonces del modelado de tópicos es descubrir cuales son los temas a los que alude un determinado conjunto de documentos (Rosati, 2020).

Se explorará la posibilidad de verificar si existe o no una modificación en la cobertura de los tópicos más relevantes de la agenda mediática en la Argentina que tiene como eje central la campaña electoral de 2015, más precisamente las elecciones Primarias Simultáneas y Obligatorias (PASO). Así mismo se describirá cuáles son las características propias de la cobertura que adquieren estos tópicos a lo largo del recorte temporal planteado.

I.2 Estado del arte

Con el propósito de analizar la cobertura mediática de las elecciones presidenciales de 2015 en la Argentina, Zunino y Marín (2016) plantean como objetivo analizar cuáles fueron los temas más recurrentes en las coberturas mediáticas de *Clarín, La Nación, Diario Uno y Los Andes*. También fijaron su atención sobre los actores involucrados en las noticias para establecer cuales fueron incluidos más asiduamente, así mismo intentaron determinar cuáles fueron las fuentes incluidas con más frecuencia en el tratamiento informativo. Los investigadores en principio corroboraron la premisa de la cobertura mediática del estilo "carrera de caballos" (Patterson, 1980)[5] y la frecuencia mayor de piezas producidas por los medios nacionales que por los medios del interior del país, "lo que constituye un indicador de las diferencias de escala de las redacciones, al mismo tiempo que corrobora que la producción de contenidos mediáticos se concentra en Buenos Aires" (Zunino y Marín, 2016:65). Así mismo, confeccionaron un listado de tópicos relevantes en las agendas mediáticas de los medios mencionados, en la cual predominaron en principio temas reunidos bajo la categoría "proselitismo", seguidos por "internas" y "economía".

La metodología elegida por los autores consistió en la construcción de un corpus de investigación compuesto por piezas periodísticas publicadas por los diarios *Clarín, La Nación, Los Andes y UNO* entre el 27 de septiembre de 2015 y el 22 de noviembre del mismo año. Es decir, desde cuatro semanas antes de la primera vuelta electoral desarrollada el 25 de octubre y hasta el día de la segunda vuelta. De esta manera el corpus quedó constituido por una población de 3155 notas. Posteriormente recurrieron a un método de muestreo de modo de reducir la cantidad de unidades de análisis a una dimensión que sea lo suficientemente pequeña como para ser estudiada y lo suficientemente amplia para ser representativa. Generaron una

[5] El término de Patterson "carrera de caballos" hace alusión a los hallazgos respecto de las coberturas mediáticas sobre periodos electorales, las cuales están dominadas por noticias sobre "ganadores y perdedores": "el contexto de juego de la elección general hacía de las perspectivas de victoria de los candidatos un tema persistente de la cobertura periodística durante toda la campaña" (como se cita en Zunino y Marín, 2016).

muestra aleatoria simple de 343 casos, los cuales constituyeron las unidades de análisis del trabajo, con un nivel de confianza del 95% y un margen de error del 5%.

Gráfico 1. Frecuencia de temas asociados a las elecciones 2015. Clarín, La Nación, Los Andes y UNO. 27 de septiembre al 22 de noviembre de 2015.

Tópico	Diario				Total
	Los Andes	Uno	Clarín	La Nación	
Proselitismo	18,80%	14,00%	16,90%	17,70%	17,10%
Internas	14,60%	18,60%	14%	16%	15,30%
Economía	22,90%	23,30%	14,00%	10,60%	15,30%
Resultados del comicio	10,40%	7,00%	16,20%	12,40%	12,90%
Debate	10,40%	11,60%	5,90%	11,50%	9,10%
Características del acto eleccionario	8,30%	4,70%	8,80%	7,10%	7,60%
Corrupción	4,20%	4,70%	5,10%	5,30%	5,00%
Justicia		4,70%	2,90%	5,30%	3,50%
Trabajo	2,10%	2,30%	2,20%	1,80%	2,10%
Seguridad / Narcotráfico	4,20%	2,30%	5,90%	1%	2,10%
Seguridad social / jubilaciones		2,30%	0,70%	2,70%	1,50%
Salud		2,30%	0,70%	1,80%	1,20%
Encuestas			0,70%	2,70%	1,20%
Educación			0,70%		0,30%
Energía			0,70%		0,30%
Otros	4,20%	2,30%	4,40%	4,40%	4,10%

Fuente: Zunino y Marín (2016). *Los medios y las elecciones: la agenda informativa de la campaña presidencial de 2015 en la Argentina.* Tabla I.

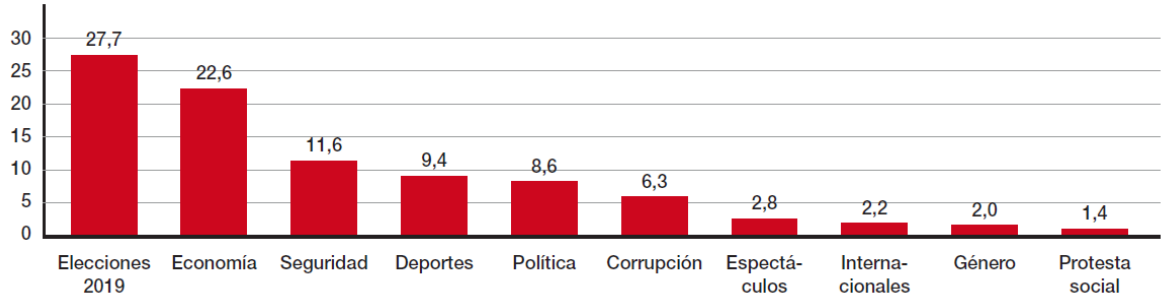
Los datos recabados no lograron confirmar las premisas planteadas por el concepto de *indexing* (Bennet, 1990)[6]. Los autores afirman que la cobertura del período analizado estuvo signada por la puesta en agenda de temas anecdóticos relacionados con los candidatos más que con sus propuestas, además de polarizar el espacio público descontextualizando y simplificando los acontecimientos de sus causales estructurales (Zunino y Marin, 2016).

Por su parte, Koziner (2020) se propone analizar el tratamiento que hicieron los principales medios de Argentina de noticias (*Clarín, La Nación e Infobae*) de los temas más importantes en el contexto de un año electoral e identificar cuáles fueron los actores acreditados como fuentes para expresarse en torno a estos temas. El análisis de contenido realizado por la autora

[6] La teoría de *indexing* afirma la existencia de una norma no escrita que, sin embargo, es incorporada de manera inconsciente por los periodistas. La misma sostiene que los trabajadores de prensa suelen establecer relaciones estables con sus fuentes, entre las se destacan las agencias gubernamentales, dado que estas exhiben un factor de autoridad que les otorga mayor verosimilitud (como se cita en Zunino y Marin, 2016).

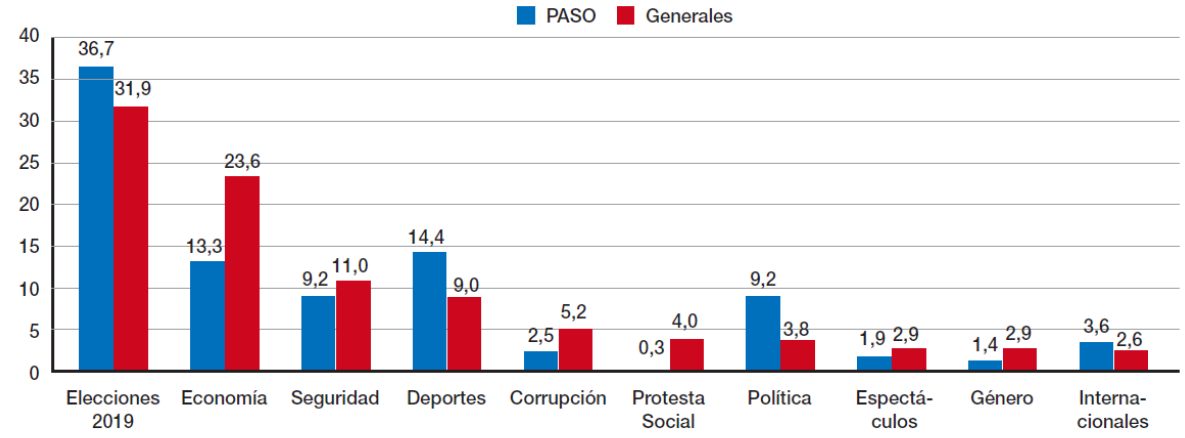
radica en la evaluación de contenido que proviene de los espacios más destacados de las páginas de inicio (o *homepage)* de los medios mencionados durante un periodo que abarca el mes de abril y octubre de 2019, de modo de incluir en el análisis ambas campañas electorales. La base empírica del trabajo fue proporcionada por el Observatorio de Medios de la Universidad de Cuyo. Los objetivos propuestos por Koziner son: 1) establecer cuáles fueron los temas más relevantes para la prensa digital entre abril y octubre de 2019, 2) identificar las fuentes de información que obtuvieron crédito en los temas más relevantes durante el período estudiado y 3) comparar el tratamiento de temas y fuentes acreditadas en las campañas para las PASO y para las elecciones generales. Para resolver los objetivos planteados la autora implementó una estrategia cuantitativa tomando las cinco primeras notas publicadas en las páginas de inicio (o *homepage)* de *Infobae, Clarín y La Nación*, en dos cortes diarios (a las 9hs y a las 19hs), de acuerdo con los momentos de mayor tráfico y actualización de las noticias, entre abril y octubre de 2019. Para la recolección se utilizó el método de una semana construida aleatoriamente por mes, de modo que el corpus quedó compuesto por 1470 noticias que constituyen las unidades de muestreo de la investigación (10 noticias diarias en tres medios durante 7 meses). Los medios analizados fueron elegidos por ser los tres más importantes del país en términos de preferencias masivas de consumo de noticias de internet, según el ranking *Alexa.com* y datos de la consultora *ComScore*. Koziner identifica que las elecciones presidenciales signaron las agendas de los principales medios digitales argentinos durante 2019, vinculándolo a la agudización de la crisis económica que atravesó el país en tal coyuntura. Corrobora también un mantenimiento constante de temáticas vinculadas a seguridad y delito a lo largo de todo el período estudiado.

Gráfico 2. Temas más importantes en Infobae, Clarín y La Nación. Período abril-octubre 2019.



Fuente: Koziner (2020). *Temas y fuentes en medios argentinos. Un estudio en contexto electoral (2019).* Gráfico 1.

Gráfico 3. Temas más importantes durante las PASO y las elecciones generales. Infobae, Clarín y La Nación. En porcentajes.

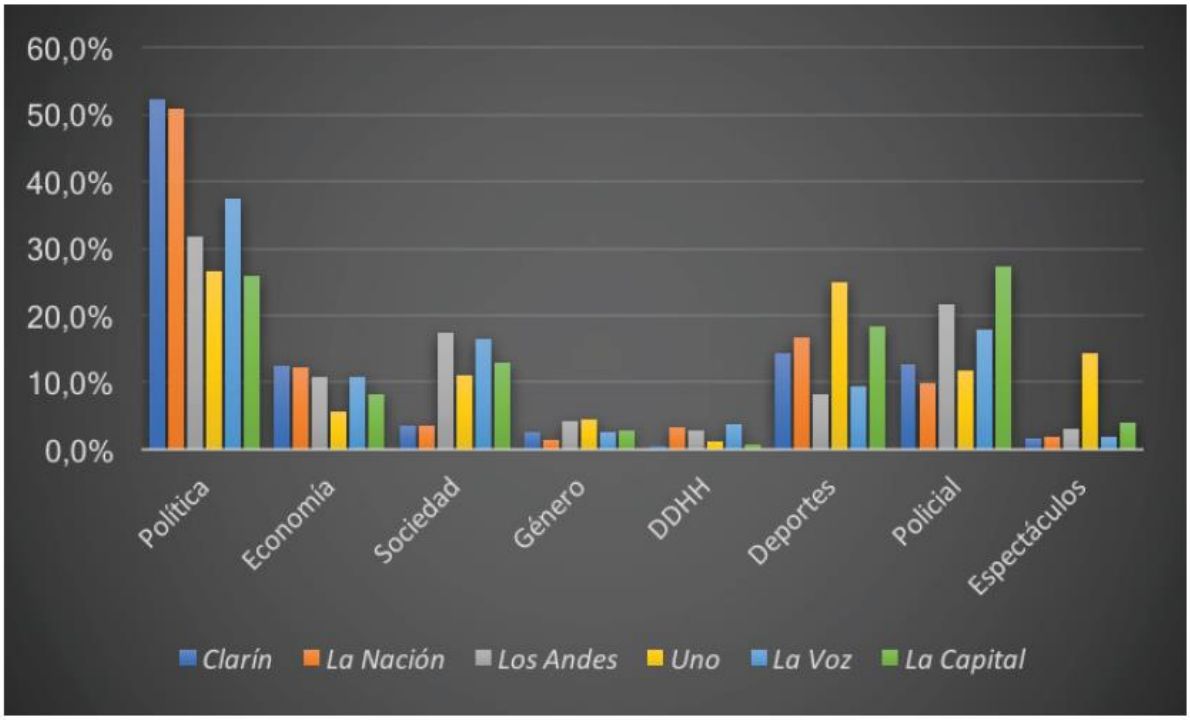


Fuente: Koziner (2020). *Temas y fuentes en medios argentinos. Un estudio en contexto electoral (2019).* Gráfico 3.

Esteban Zunino y Augusto Grilli Fox (2019) también indagan sobre medios digitales en la Argentina. El trabajo citado tiene como objetivo analizar los contenidos mediáticos de seis medios *online* de las ciudades más importantes de la Argentina con el fin de evaluar de qué modo la potencialidad de los medios digitales redundó en mayor calidad informativa. La metodología cuantitativa fue aplicada sobre un corpus de 3360 noticias que constituye las unidades de análisis de investigación del trabajo. Se codificaron en tiempo real en dos cortes diarios coincidentes con momentos de alta actualización y tráfico (09:00 y 19:00 Hs), en cuatro cortes temporales de dos semanas cada uno a lo largo de 2017 y 2018. En cada uno de los cortes

seleccionados se codificaron las primeras cinco noticias de las páginas de inicio (*homepage)*, considerando que la ubicación es un criterio clásico de jerarquía informativa que los medios digitales utilizan. Se realizó un trabajo de intercodificación de 336 noticias que corresponden al 10% del corpus. La elección de las noticias de la muestra de intercodificación surge de una estrategia de estratificación: se seleccionan 84 noticias de cada una de las cuatro etapas de análisis teniendo en cuenta que 14 fueran de cada uno de los diarios seleccionados. La selección de las 14 piezas correspondientes a cada medio se extrajo de manera aleatoria. Los autores concluyen la imposición de temas políticos y económicos durante el período estudiado, aunque también señalan la preponderancia de temas blandos como policiales y deportes.

Gráfico 4. Frecuencia de temas presentes en los medios online. Clarín, La Nación, Los Andes, UNO, La Voz y La Capital. 2017-2018.



Fuente: Zunino E y Grilli Fox A (2019). *Medios digitales en la Argentina: posibilidades y límites en tensión.* Gráfico 10.

Aruguete y Calvo (2018) plantean que los mensajes se propagan con distinta velocidad en la red y que la diferencia en la propagación depende de la congruencia o disonancia cognitiva existente entre el usuario y el contenido de los mensajes publicados. Los usuarios activan contenido que componen agendas colectivas (*agenda melding*), insertando vínculos a medios tradicionales y no tradicionales, limitando la capacidad de los medios masivos a fijar la agenda pública. Otros trabajos que se inscriben en esta propuesta son los de Ernesto Calvo (2015) o

Paulo-Carlos López-López y Javier Vásquez-González (2018), quienes abordan la cuestión de la llamada "*social media issue agenda*" analizando Twitter. Tales trabajos indagan sobre la irrupción de las herramientas desarrolladas en esta sociedad en red, que además de favorecer una fragmentación de las audiencias, promueven un cierto cambio metodológico en la investigación de la agenda que se configura como aquella que discuten los ciudadanos en este nuevo espacio público y que reconfigura la tradicional visión sobre la preminencia de la agenda política y la agenda mediática (como se cita en Paulo-Carlos López-López y Javier Vásquez-González, 2018). El establecimiento de los temas relevantes en la esfera pública en la era de los medios 2.0 es aún un campo en expansión, sin embargo, no existe un consenso uniforme respecto a los mecanismos que logran configurarla. La indagación sobre la relación entre los medios tradicionales y los espacios virtuales es reciente y, por cierto, asistemática, por lo que distintos investigadores vienen reclamando mayor atención en ese ámbito (Kushin, 2010; Meraz, 2009, 2011). En efecto, no hay una respuesta unívoca que indique si los nuevos medios disputan la agenda de los medios tradicionales o si repiten su temario, menos aún si tienen la capacidad de establecer la agenda pública (Aruguete, 2015).

El presente trabajo, si bien recupera interrogantes cercanos a los de los autores mencionados, intenta abordarlos a partir de la aplicación una metodología proveniente del campo del aprendizaje automático. Para esto, se analizarán las noticias de los grandes medios de comunicación a nivel nacional en el periodo comprendido entre enero de 2015 y diciembre de 2016, utilizando el modelo generativo *LDA (Latent Dirichlet Allocation)*, para detectar tópicos latentes representados en clúster de palabras, basado en un modelo estadístico del lenguaje. Es posible destacar dos puntos donde la aplicación de técnicas computacionales ofrece mejoras. Por un lado, mencionar el factor de la escalabilidad en el análisis de las noticias. Los textos expuestos en el estado del arte analizan entre 1400 y 3600 noticias en total, siendo que en el presente trabajo el corpus de texto está compuesto por 466.754 noticias. Por otro lado la posibilidad de replicabilidad del proceso permite explicitar claramente el proceso de extracción de valor de los datos de manera completa, desde su recolección hasta su análisis.

Replicabilidad, escalabilidad, sistematización y cierto grado de automatización, aplicación de procesamiento de lenguaje natural y otras técnicas cuantitativas basadas en aprendizaje automático o redes neuronales son posibles gracias al desarrollo y proliferación del campo disciplinar de las ciencias sociales computacionales. En el siguiente apartado se realizará un breve recorrido histórico sobre el desarrollo de esta disciplina, en la cual se enmarca el presente trabajo. Posteriormente se discutirá la metodología escogida y su implementación.

I.3 Ciencias sociales computacionales

Tal como expone Cioffi-Revilla (2010) las ciencias sociales computacionales como disciplina data de la segunda mitad del siglo 20, junto con la invención de las computadoras electrónicas. Durante la década de 1960 y 1970 los científicos sociales empezaron a utilizar computadoras para conducir análisis estadísticos, esos fueron los días pioneros de SPSS, SAS y los trabajos con tarjetas perforadas *(punch-cards)*. Los fundadores de una orientación más teórica en las ciencias sociales computacionales durante la primera generación incluyen a Herbert A. Simon (1916-2001), Karl W. Deutsch (1912-1992), Harold Guetzkow (1915-2008) y Thomas C. Schelling (1921-2016). Las ciencias sociales computacionales son una disciplina científica *instrument-enabled*[7] similares a la microbiología, la radioastronomía o las nanociencias – nuevos campos científicos posibles de existir gracias al microscopio, el radar y al microscopio electrónico respectivamente. En cada una de estas disciplinas, incluyendo las ciencias sociales computacionales, es el instrumento de investigación el que conduce el desarrollo de la teoría y el entendimiento.

El surgimiento de técnicas provenientes de las ciencias de la computación con la popularización de la *big data* suscitó en las ciencias sociales computacionales una serie de discusiones de las cuales se intentará dar cuenta. Se abordará brevemente la historia del conjunto de técnicas denominada aprendizaje automático, desde conceptos fundacionales hasta el estado actual de situación. A la par de las ventajas del uso de técnicas computacionales, surgen también desafíos, límites y problemas en el proceso de su implementación en investigación de ciencias sociales.

Warren Weaver (1984) escribió un influyente artículo sobre ciencia y complejidad en la revista *American Scientist* en el que explica básicamente que existieron tres tipos de problemas en los que estuvo trabajando la historia de las ciencias. En periodos previos a 1900, la ciencia operaba sobre lo que Warren denomina "problemas de simplicidad", o problemas de dos variables. Por ejemplo, en física, temperatura y presión o en las ciencias sociales, población y tiempo y su correlación con otras variables como la producción o el comercio. Posterior a 1900 los "problemas de promedios" toman relevancia y se menciona cómo los científicos desarrollaron poderosas técnicas basados en la teoría de la probabilidad para enfrentar problemas donde el número de variables es muy grande, y problemas donde cada una de las diversas variables tiene

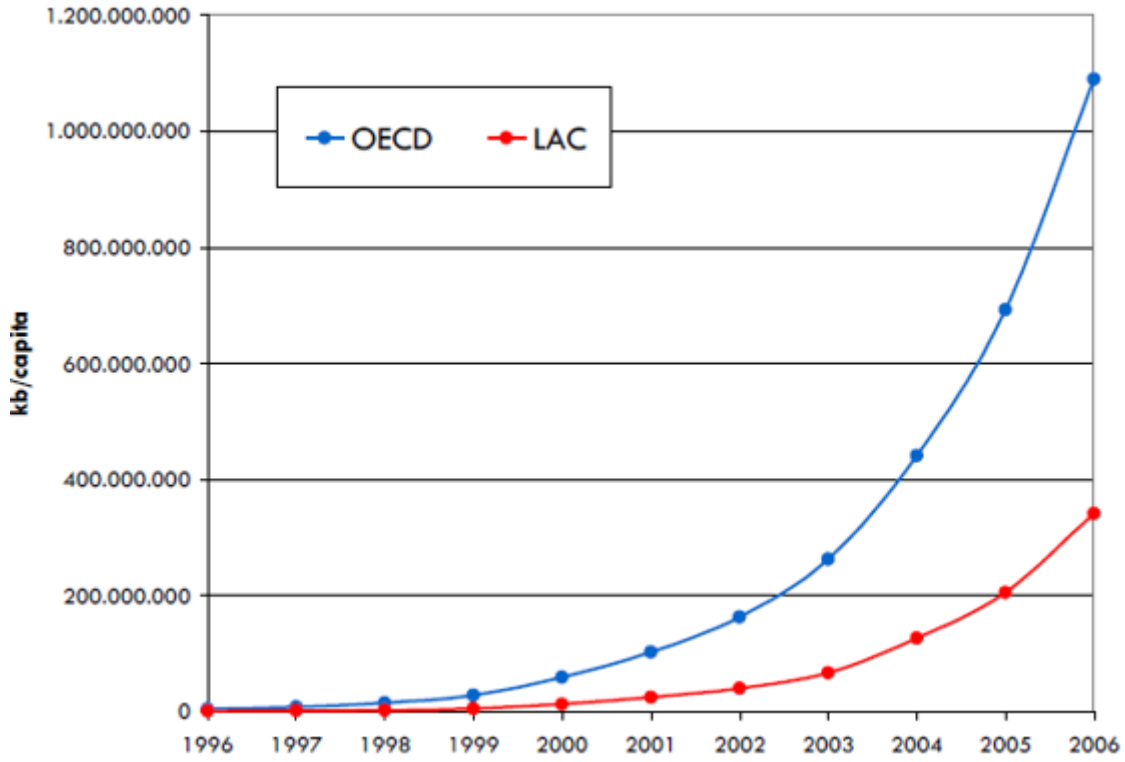
[7] Por *instrument-enabled* se entiende literalmente "permitidas por el instrumento", en este caso haciendo referencia a la posibilidad de existencia de una disciplina científica por las posibilidades que otorga determinado instrumento, en este caso, la computadora electrónica.

un comportamiento quizá totalmente impredecible. El último grupo se denomina de "complejidad organizada" donde, utilizando simultáneamente un número considerable de factores que están interrelacionados en un todo orgánico, los problemas de disciplinas como la economía, las ciencias políticas y la sociología no pueden ser tratados con técnicas estadísticas efectivas en el marco de "problemas de promedios". Estos nuevos problemas, requieren a las ciencias realizar un gran tercer avance que debe ser más grande que la conquista de los problemas de simplicidad del siglo XIX o la de los problemas de "complejidad desorganizada" del SXX. La ciencia debe, explica el autor, en los próximos 50 años, aprender a enfrentar estos problemas de "complejidad organizada", agregando que el avance de los dispositivos computacionales tendrá un gran impacto en la ciencia, *"they will make it possible to deal with problems wich previously were too complicated, and, more importantly, they will justify and inspire the development of new methods of analysis applicable to these new problems*"[8] (Weaver, 1948:541).

Las ciencias sociales computacionales pretenden abordar el estudio de los llamados fenómenos de "complejidad organizada". El hecho fundante que devino en la popularización y revitalización de la subdisciplina de las ciencias sociales computacionales es la llamada "revolución digital". Tanto la disponibilidad de la información digital, como la capacidad de almacenamiento y procesamiento de los equipos informáticos ha aumentado de manera exponencial teniendo como clivaje el año 2000 – 2001 (Hilbert & Lopez, 2011).

[8] "… (los dispositivos computacionales) harán posible enfrentar problemas que previamente eran muy complicados y, aún más importante, justificaran e inspiraran el desarrollo de nuevos métodos de análisis aplicables a estos nuevos problemas " (traducción propia).

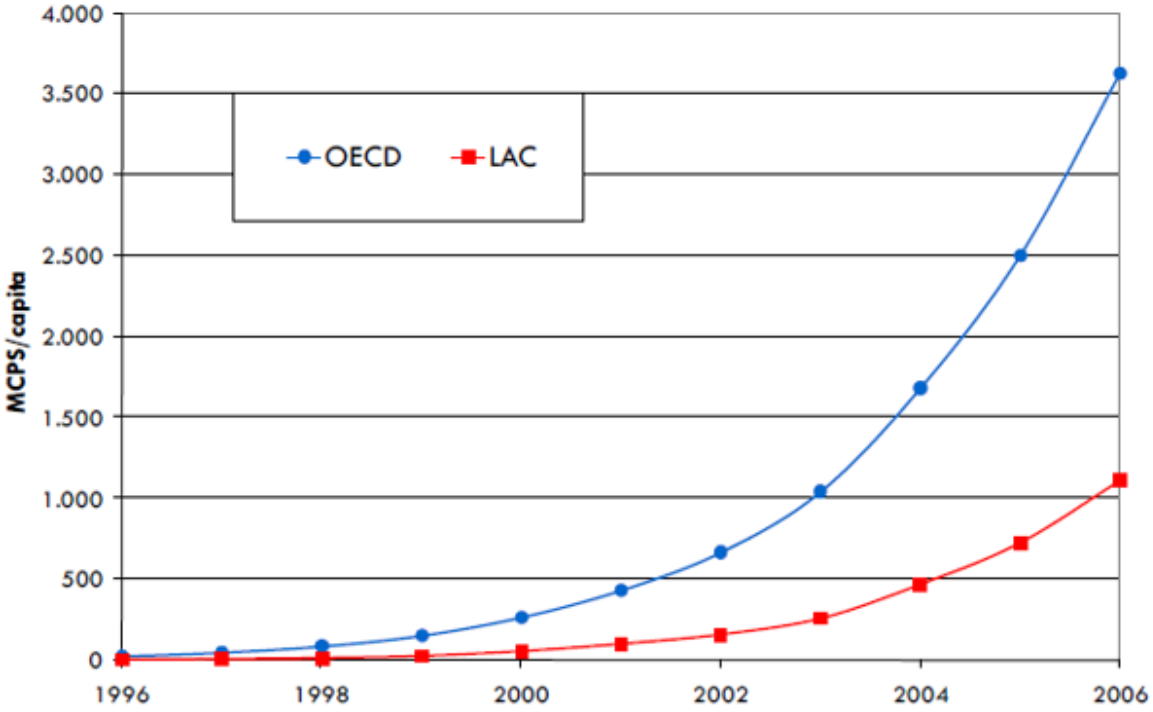
Gráfico 5. Capacidad de almacenamiento de información en discos duros de computadores personales y computadoras portátiles.[9]



Fuente: Hilbert & Lopez (2011). *The World's Technological Capacity to Store, Communicate, and Compute Information*. Figure 16.

[9] El eje Y expresa el valor kb per capita. La unidad referida es el kilobit.

Gráfico 6. Capacidad de cómputo de información con computadores personales, computadoras portátiles y teléfonos celulares[10].



Fuente: Hilbert & Lopez (2011). *The World's Technological Capacity to Store, Communicate, and Compute Information.* Figure 17.

La disponibilidad de recursos informáticos son parte esencial del desarrollo de las ciencias sociales computacionales. Para ilustrar, en la época en la escribía Warren Weaver sobre los problemas de complejidad organizada, un disco de apenas 5 megabytes pesaba alrededor de una tonelada y su costo de operación lo tornaba muy poco accesible.

[10] El eje Y del gráfico muestra el valor MCPS per capita. MCps o *Million Computations per second* es una medida de capacidad de cómputo de uno o varios procesadores. Expresa millones de cómputos por segundo.

Fotografía 1. Transporte de disco duro IBM de 5 megabytes (1956).

Fuente: Historical Pictures Inc.

Desde mediados de los años 1990 se ha expandido de modo sistemático el uso intensivo de métodos computacionales para el estudio de procesos sociales en antropología, economía sociología, arqueología y ciencia política. Sin embargo, la modelización computacional como metodología de investigación científica constituyen un enfoque poco empleado en la actualidad por parte de las ciencias sociales latinoamericanas. El concepto de modelo no forma parte del

habitus metodológico de las ciencias sociales por cuestiones vinculadas al propio desarrollo histórico de las ciencias y sus disciplinas (Rodriguez Zoya y Roggero, 2014). El uso del concepto de modelización como eje ordenador constituye una estrategia eficiente al momento de abordar la complejidad social. El trabajo de modelizar permite una verdadera zona de intercambio (McFarland, Lewis y Goldberg, 2015) para construir conocimiento interdisciplinar en diálogo con las ciencias de la vida, las ciencias de la materia y las ciencias computacionales.

I.3.1 Antecedentes y actualidad del aprendizaje automático

Alan Turing exploró las posibilidades matemáticas de la inteligencia artificial en su escrito "*Computing machinery and inteligence*" (1950), donde sugiere que, si los humanos usan información disponible y el raciocinio para resolver problemas y tomar decisiones, ¿por qué no podría una máquina? En este marco lógico, Turing discute la idea de construir máquinas inteligentes y cómo evaluar su inteligencia. ¿Qué se interpuso entre Turing y la construcción y evaluación de estas máquinas inteligentes? Como señalábamos previamente, la disponibilidad de recursos computacionales era limitada. Previo a 1949, la computadora carecía del prerrequisito necesario para la inteligencia: la memoria. Las computadoras no podían almacenar comandos, solo los podían ejecutar; en otras palabras, se le podía decir a las computadoras qué hacer pero ellas no recordaban qué es lo que hacían. A su vez, los costos eran muy elevados. Solamente prestigiosas universidades y grandes compañías tecnológicas podían asumir tales gastos.

Cinco años después surgía el programa *Logic Theorist* de Allen Newell, Cliff Shaw y Herbert Simon, el cual es considerado uno de los primeros programas de inteligencia artificial, ya que se proponía simular las habilidades de resolución de problemas de una persona. Fue presentado al público en 1956 en la *Darthmouth Summer Research Proyect on Artificial Inteligence*, uno de los eventos catalizadores del campo de la inteligencia artificial (Anyoha, 2017). De 1957 a fines 1980, las computadoras podían guardar más información y se volvieron más rápidas, baratas y accesibles. Los algoritmos de aprendizaje automático mejoraron, sumado a que las personas mejoraron en función de definir qué algoritmo aplicar para resolver determinado problema. Al respecto, Carbonell, Michalski y Mitchell (1983) reconstruyen la historia del aprendizaje automático y explican con precisión los antecedentes históricos, separando entre tres grandes periodos centrados alrededor de diferentes paradigmas: a) modelización neuronal y los modelos teóricos para la toma de decisiones, b) aprendizaje simbólico orientado por

conceptos y c) aproximaciones conocimiento-intensivas combinando varias estrategias de aprendizaje.

Las primeras redes neuronales datan de los inicios del aprendizaje automático, dada la primitiva naturaleza de la tecnología computacional las investigaciones bajo este paradigma eran teóricas o involucraban construcciones ad-hoc de sistemas de hardware como *perceptrons* (1958), *pandemónium* (1959) y *adelaine* (1962). Los paradigmas simbólicos orientada por conceptos y varios sistemas de reconocimiento de patrones datan de 1963 a 1974. También destacan programas basados en aprendizaje inductivo y trabajos relacionados. El tercer paradigma nominado conocimiento-intensivo moderno, es el más reciente en términos históricos. Encuentra sus inicios durante la mitad de los '70 y destacan trabajos como la exploración de métodos alternativos de aprendizaje, el aprendizaje por analogía y el descubrimiento de conceptos y clasificaciones[11].

Durante la década de 1980 surgió el segundo atolladero, las computadoras podían guardar comandos y ejecutarlos, pero su potencia era limitada y para emular la capacidad de raciocinio de resolución de problemas de las personas (por ejemplo emular un enunciado en respuesta a otro), la computadora precisa saber todo el lenguaje, necesita conocer combinaciones, realizar procesamientos en simultáneo. La capacidad de procesamiento funcionó a modo de cuello de botella, lo que recién en la década 1990-2000 va a comenzar a resolverse de manera positiva, y la inteligencia artificial comenzaría a entregar verdaderas pruebas de realidad, teniendo como corolario el famoso partido de ajedrez que Gary Kasparov perdió en 1997 jugando contra *Deep Blue*, una computadora programada para tal fin de la compañía IBM (Anyoha, 2017). Por esos años también era implementado el primer software de reconocimiento de voz (desarrollado por *Dragon System*, implementado en Windows) y el primer robot que podía reconocer y expresar emociones, *Kismet*, desarrollado por Cynthia Breazeal. El paradigma dominante en la inteligencia artificial de 1950 a 1980 es la inteligencia artificial simbólica[12]. En la década de 1990 la subdisciplina de la inteligencia artificial genera algunos grupos de intereses diversos, uno de ellos es el campo del aprendizaje automático o *machine learning* el cual cobra una especial popularidad. El objetivo central pasó de crear una inteligencia artificial a resolver

[11] Para información en detalle sobre autores y trabajos pioneros consultar Carbonell, Michalski y Mitchell (1983).

[12] La inteligencia artificial simbólica usa símbolos interpretables por humanos que representan entidades del mundo real o conceptos en función de crear 'reglas' para la manipulación concreta de esos símbolos, creando sistemas basados en reglas. En resumen, la IA simbólica implica la incorporación explicita de reglas de comportamiento y conocimiento humano en programas de computación.

problemas de naturaleza práctica, se corrió el foco de los abordajes simbólicos que había heredado de la inteligencia artificial, para pasar a métodos y modelos prestados de la estadística y la teoría de la probabilidad. El aprendizaje automático tiende a utilizar grandes y complejos set de datos por lo que los análisis estadísticos clásicos como los análisis bayesianos serían impracticables. Como resultado, el aprendizaje automático y especialmente su subconjunto de técnicas más populares en la actualidad conocidas como aprendizaje profundo o *deep learning*, exhiben comparativamente una orientación focalizada a la ingeniería. Es una disciplina práctica donde las ideas son testeadas empíricamente antes que teóricamente (Chollet, 2018).

I.3.2 El aprendizaje automático como estrategia metodológica en investigaciones de ciencias sociales

Existen enfoques epistemológicos con posiciones extremas[13] que plantean (más o menos explícitamente) un retorno a un empirismo ingenuo. Ven en la alta disponibilidad de datos y recursos computacionales la posibilidad de aplicar un gran número de algoritmos a un set de datos para determinar cuál es el mejor, o generar un modelo compuesto o explicación, lo que permitiría prescindir de la teoría y del conocimiento de dominio específico. Quienes defienden la utilización de estos enfoques, explican que la percepción del mundo se construye desde un abordaje diferente; en lugar de testear una teoría analizando datos relevantes, las nuevas técnicas analíticas de los datos permiten obtener información valiosa "nacida desde la información". Plantean la existencia de un nuevo paradigma en la ciencia, llamado "ciencia exploratoria" donde predominan los enfoques data-intensivos, la exploración estadística y la minería de datos. Referentes de este nuevo paradigma afirman que el aluvión de datos hace el método científico obsoleto, ya que los patrones y relaciones que nacen de la *big data* produce de manera inherente resultados significativos sobre fenómenos complejos (Kitchin, 2014). Robert Kitchin polemiza con aquellos quienes reivindican estos enfoques radicalizados proponiendo la combinación de enfoques conducidos por teoría y enfoques conducidos por datos, tomando lo mejor de las dos metodologías.

Desde la estadística, Leo Breiman (2001) advierte sobre el uso exclusivo del modelado tradicional de datos basado en supuestos fuertes, lo que ha llevado a teoría irrelevante o

[13] Por ejemplo, Anderson (2008) afirmó "Ahora hay un mejor camino. Los Petabytes nos permiten decir "correlación es suficiente" … Podemos analizar los datos sin hipótesis sobre qué podría mostrar. Podemos tirar números en los clústeres de computadora más grande que el mundo ha visto y dejar que los algoritmos estadísticos encuentren patrones donde la ciencia no puede… La correlación reemplaza a la causalidad, y la ciencia puede avanzar aún sin modelos coherentes, teorías unificadas o realmente cualquier mecanismo de explicación. No hay razón para aferrarse a los viejos métodos."

tautológica, conclusiones cuestionables y ha mantenido a los profesionales de la estadística alejados de la posibilidad de trabajar en un largo rango de problemas actuales interesantes. Aboga por la incorporación del modelado algorítmico, tanto en la teoría como en la práctica, dado que puede usarse en grandes y complejos set de datos de manera más eficaz e informativa que los modelos de datos tradicionales en set de datos más pequeños. Insta a retirar a la disciplina estadística de la dependencia exclusiva de modelos de datos tradicionales y adoptar set de herramientas más diversas que ofrece la ciencia de datos y la aplicación de técnicas computacionales.

Ruth Sautú (2019) advierte que '*big data'* implica no solo un cambio en la fuente y construcción de los datos y su volumen, sino también replantea los supuestos ontológicos-epistemológicos que sustentan la investigación:

> "En muchas bases ignoramos la concepción de la realidad y las maneras como han sido definidas las unidades originales de las que los datos emanan. Más aún, nos posicionamos en el nivel del dato que estadísticamente nos muestra la presencia de secuencias y asociaciones, pero no necesariamente llegamos a comprender qué procesos individuales y sociales subyacen a ella. En la investigación científica es clave tener en cuenta cómo han sido producidos originalmente los datos, cómo han sido recogidos y registrados, y tener la posibilidad de establecer la consistencia interna de la base de datos. (…) La clave de la validez de un análisis secundario de datos reside en el razonamiento sobre el referente teórico de las variables; es decir, su definición en el marco de una teoría." (Sautú, 2019:108)

De los planteos críticos a las defensas radicales surgen algunas respuestas que intentan generar una instancia intermedia de combinación de enfoques que integren las estrategias que provienen del aprendizaje automático o el análisis estadístico computacional combinando elementos de los enfoques guiados por teoría social. Rodriguez Zoya y Roggero (2019) explica que los métodos computacionales y los métodos clásicos afirman su carácter de complementariedad mutua o una vía de triangulación metodológica, dado que permite integrar distintos tipos de evidencia empírica de tipo cualitativo y cuantitativo. McFarland, Lewis y Goldberg (2015) plantean la posibilidad de la *big data* como una zona de intercambio y postula la idea de las "ciencias sociales forenses" en un esfuerzo por generar un enfoque híbrido que combina perspectivas aplicadas y conducida por datos integrando abordajes inductivos y deductivos como mutuamente informativos, usando la teoría para guiar parcialmente la exploración deductiva de los datos mientras el enfoque deductivo nos ayuda a descubrir qué teoría nos proveen de explicaciones. Kitchin (2014) caracteriza a la "ciencia conducida por datos" como un enfoque híbrido que combina abordajes abductivos, inductivos y deductivos

para comprender un fenómeno determinado. El presente trabajo intenta ser una implementación de un enfoque hibrido, en el sentido que McFarland, Lewis y Goldberd (2015) le dan al término. La teoría juega un rol clave en la demarcación del objeto de estudio que se propone, combinando metodologías data – intensivas para evaluar la hipótesis propuesta.

I.4 Limites y alcances del presente trabajo

El actual trabajo toma nociones fundamentales de los autores McCombs, Eilders y Aruguete, aunque no se pretende realizar un debate con las teorías que se ocupan del establecimiento de la agenda. Más bien se intentará responder a ciertos interrogantes clásicos que se desprenden de tales posiciones teóricas apelando a una metodología que proviene del campo del aprendizaje automático, enmarcada en las denominadas ciencias sociales computacionales. Para esto, se analizarán las noticias de los grandes medios de comunicación *online* a nivel nacional en el periodo comprendido entre enero de 2015 y diciembre de 2016, utilizando el modelo generativo LDA (*Latent Dirichlet Allocation*), para detectar tópicos latentes representados en clúster de palabras, basado en un modelo estadístico del lenguaje. El modelo generativo LDA es un modelo de aprendizaje automático no supervisado, esto quiere decir que no hay explícitamente una variable a predecir. No hay una variable dependiente. Lo que hay es una gran cantidad de variables independientes (en este caso las palabras) a las que el modelo se ajusta. El modelo busca patrones previamente no detectados en un set de datos sin etiquetas preexistentes y con mínima supervisión humana.

En el presente trabajo un concepto importante que se menciona desde el título es "tópico". Se optó por definir al "tópico" en un sentido estrictamente técnico: un tópico será una distribución de probabilidad a lo largo de un vocabulario de un corpus determinado. De manera general, los supuestos clave del modelo LDA son que un documento (en el caso de este trabajo, una noticia) está compuesto por varios tópicos. Esto quiere decir que, por ejemplo, una misma noticia sobre política general puede incluir entre sus temas menciones a hechos de inseguridad, de política internacional y cuestiones proselitistas, las cuales, si bien existen, no son la distribución predominante. El tópico está compuesto de palabras, específicamente, es una distribución de probabilidad a lo largo de un vocabulario y preexisten a los documentos (Mutzel, 2015). Lo anteriormente expuesto es una aproximación a una operacionalización técnica de la definición conceptual de tópico de Pan y Kosicki (1993) expuesta en el marco teórico[14].

[14] Ver página 5.

Capítulo II: Metodología

II.1 Una introducción al procesamiento de lenguaje natural

Bird, Klein y Loper (2009) definen lenguaje natural como el usado todos los días para comunicarnos entre seres humanos. Lenguajes como inglés, español o portugués. En contraste a los lenguajes artificiales (como por ejemplo el código de un software), el lenguaje natural evoluciona mientras se pasa de generación en generación y no es simple codificarlo con reglas explícitas.

Como plantean los autores mencionados, se comprende al procesamiento de lenguaje natural en un sentido amplio, abarcando cualquier tipo de manipulación computacional del lenguaje natural. Contempla desde un simple conteo de frecuencia de palabra para evaluar distintos estilos de escritura hasta herramientas que buscan aproximarse a la semántica de una enunciación humana, al menos en la medida que permite la compresión y la generación de una respuesta útil a tal enunciación. Las tecnologías vinculadas al procesamiento de lenguaje natural son cada vez más difundidas y utilizadas en diferentes investigaciones. Por ejemplo, teléfono y computadoras de bolsillo que usan texto predictivo y reconocimiento de escritura manual, traducción por computadoras que permite enviar un texto escrito originalmente en chino mandarín a un recipiente en idioma español. Proveyendo interfaces humano-máquina más naturales y accesos más sofisticados a información almacenada, el procesamiento de lenguaje natural juega hoy en día un rol fundamental en la sociedad multilingüe de la información. Los análisis utilizando *NLP (Natural Language Processing)* en investigaciones científicas son un campo fértil y de fuerte desarrollo en el último siglo entre los que destacan los trabajos de Garg *et al.* (2018) sobre estudios de estereotipos étnicos y de género, Wang *et al.* (2012) sobre análisis de sentimiento en tiempo real sobre la red social *Twitter* y Gálvez *et al.* (2019) quienes indagan sobre asociaciones estereotípicas entre género e intelecto en películas. Es cierto que el procesamiento de lenguaje natural en algunas subdisciplinas de la academia se lo conoce como "lingüística computacional", sin embargo, no se emplea tal término en el presente trabajo.

En una técnica de análisis no supervisado[15] como es el modelo *Latent Dirichlet Allocation (LDA)* de detección de tópicos, no hay una variable determinada a predecir, lo que hay en este

[15] En un modelo de aprendizaje supervisado el algoritmo aprende de un set de datos etiquetado, es decir, se le provee una etiqueta con la respuesta correcta la cual el algoritmo utiliza para evaluar su precisión sobre el set de datos de entrenamiento. En un modelo no supervisado, en contraste, los datos no tienen ninguna etiqueta, por lo que el algoritmo intenta darle sentido a los datos extrayendo patrones y atributos por sí solo.

caso son variables independientes (las palabras). Una vez que se recolecta toda la información a utilizar (algo que se abordará en la estrategia metodológica), el inicio del preprocesamiento de los datos se da al armar la matriz de término-documento. Se denomina documento a cada uno de los artículos de noticia que forman parte de la base de datos, y se denomina corpus a la colección de todos los documentos. El corpus de texto que se trabaja en el presente estudio consta de 466.754 artículos de noticias. En otras palabras, el corpus de texto consta de 466.754 documentos.

Gráfico 7. Representación ilustrativa de un corpus de texto.

Fuente: Elaboración propia en base a Rosati (2019).

La base de datos entonces comprenderá de un registro u observación por cada documento, donde se almacenará el enlace del artículo, la fecha de publicación, su título y su cuerpo.

Cuadro 1. Primeras cinco entradas de la base de datos final.

	titulo	texto	fecha	Link
0	Tinelli, a punto de venderle sus acciones a López	De dueño a empleado. Ese el recorrido que hará...	2016-04-10 02:45:00	http://www.clarin.com/extrashow/tv/Tinelli-pun...
1	Los bonos de la emisión no tendrán calificació...	El road show para salir a buscar al mercado un...	2016-04-10 02:45:00	http://www.ieco.clarin.com/bonos-emision-calif...
2	Solteras a los 27, el drama de las "mujeres so...	Las mujeres solteras mayores de 27 años en Chi...	2016-04-10 02:45:00	http://www.lanacion.com.ar/1887974-solteras-a-...
3	Santa Cruz: Gendarmería custodia una escribaní...	Una patrulla de Gendarmería Nacional custodia ...	2016-04-10 02:30:00	http://www.telam.com.ar/notas/201604/142737-la...
4	Defensa y Justicia derrotó a Temperley en Flor...	El partido que hace unos años hubiese corresp...	2016-04-10 02:00:00	http://www.infobae.com/2016/04/09/1803217-defe...

Fuente: Elaboración propia sobre información de GDELT.

Como se puede apreciar en el cuadro 1, cada registro o fila es una noticia, y cada columna contiene una porción específica de información de la noticia como lo es el título, el texto (bajada y texto principal), la fecha y el enlace que nos remite al artículo almacenado. El set de datos necesario para realizar la modelización propuesta requiere de cierta especificidad en la información que contiene. El análisis se realiza sobre palabras que sean informativas sobre el contenido de tal documento, esto en otras palabras quiere decir que es necesario "limpiar" o quitarle a la base de datos términos que cargan con poca información, como por ejemplos preposiciones, pronombres, artículos, diferentes tipos de conectores, y demás texto que se considere apropiado borrar por comprenderlo como no informativo. La palabra "donde" probablemente no aporte tanta información a un tópico determinado como sí lo haga la palabra "inflación" o "elecciones". La poca información que aporte un término puede estar dado tanto por su excesiva repetición como casi nula presencia en el corpus.

Una vez finalizada esta tarea, comienza la confección de lo que se conoce como una matriz de frecuencia término-documento. Esto implica la generación de una tabla donde cada fila representa un término y cada columna representa un documento del corpus de texto. Lo que se realiza entonces es un conteo de frecuencia, es decir cuántas veces aparece la palabra en cada uno de los documentos.

Cuadro 2. Matriz de frecuencia término-documento.

	Documento 1	Documento 2	Documento 3	Documento 4	Documento 5	...
nisman	1				1	...
fiscal	2		2			...
congreso	1	1			2	...
suicidio	1		1			...
...	...	...	...	...	...	...

Fuente: Elaboración propia en base a Rosati (2020).

Los inconvenientes que surgen al realizar esta operación sobre los datos son varios. En principio se pierde la secuencia de ordenamiento original de las palabras en la noticia. Una forma de recuperar parcialmente el ordenamiento en función del análisis es implementar bigramas o trigramas, esto es, conjunto de palabras (dos, tres, o más) que siempre se dan en coocurrencia sucesiva como por ejemplo bigramas como “Buenos Aires” o “San Lorenzo” como así también trigramas como “lucha de clases” si es un texto académico o “Newell’s Old Boys” si hablamos de deportes.

Por otro lado, se debe tener en cuenta que se está haciendo un conteo absoluto de las palabras, un primer problema que puede aparecer es, por ejemplo, la desigual extensión de los textos que se puede cargar al modelo. Si en un mismo corpus existe un documento de 400 palabras y otro de 5000, la frecuencia que refleje la matriz seguramente se verá afectada por tal motivo.

Cuadro 3. Matriz de frecuencia término-documento de conteo absoluto a distribución de una proporción.

	Documento 1	Documento 2	Documento 3	Documento 4	Documento 5	...
nisman	1				1	...
fiscal	2		2			...
congreso	1	1			2	...
suicidio	1		1			...
...	...	...	...	...	...	...

	Documento 1	Documento 2	Documento 3	Documento 4	Documento 5	...
nisman	0,5				0,5	...
fiscal	0,5		0,5			...
congreso	0,25	0,25			0,5	...
suicidio	0,5		0,5			...
...	...	...	...	...	...	...

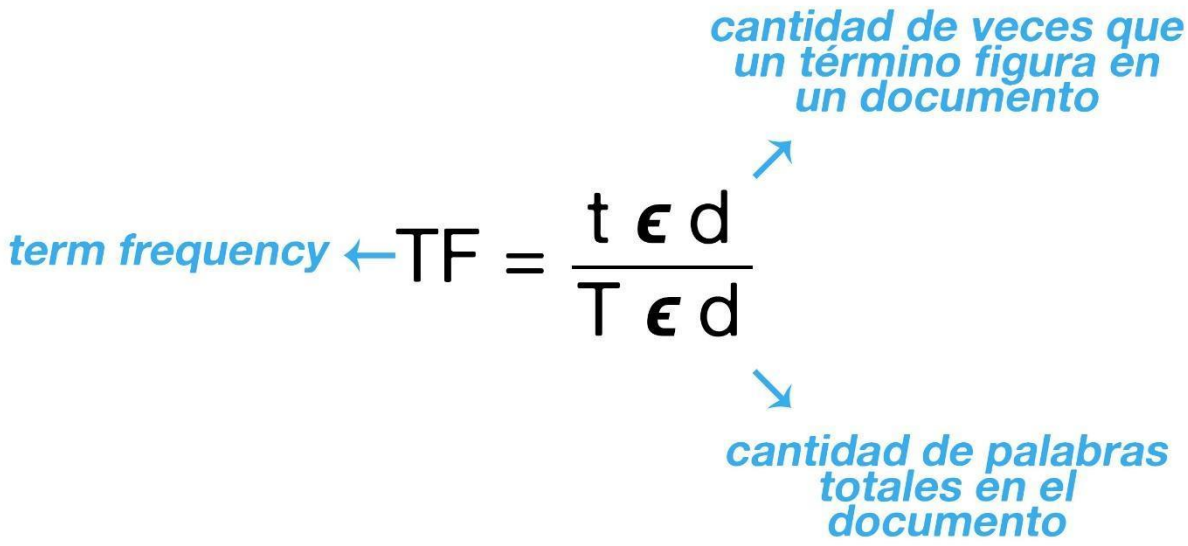
Fuente: Elaboración propia en base a Rosati (2020).

Como explica Rosati (2020) respecto a la ponderación de las palabras, en función de definir qué término aporta más información a los fines del modelado de tópicos, podemos pensar en dos dimensiones de las frecuencias de los términos de un corpus:

1- Un término *t* es más **importante** si es más frecuente en un documento *d* de un corpus *C* determinado
2- A su vez, *t* es más **informativo** del contenido de un documento *d* si está presente en pocos documentos y no en todos de *C*.

Es decir, es necesario considerar tanto la frecuencia de *t* a lo largo de todo el corpus *C* y al interior de *d*. Existen varias opciones para abordar esta problemática. Respecto a la medición de la importancia de un término en un documento, se utilizará la fórmula de frecuencia de término o *term frequency*. Esta operación devuelve el peso del término dentro del documento considerando el corpus suministrado.

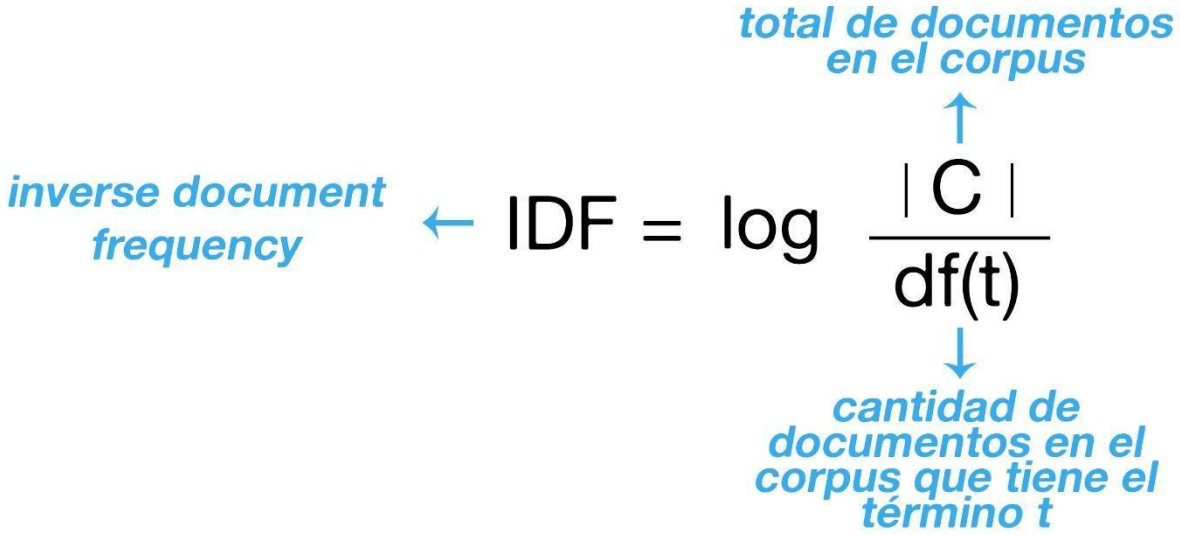
Gráfico 8. Fórmula de frecuencia de término.



Fuente: Elaboración propia en base a Rosati (2019).

Para medir la informatividad de un término se emplea la métrica de la inversa de frecuencia de documento o *inverse document frequency*, esto es, sobre el total de documentos se calcula en cuantos documentos aparece el término. Se utiliza la inversa de la frecuencia de documentos o *document frequency* dado que permite una lectura más intuitiva, mientras más grande sea IDF, menor será el grado de repetición de término en los documentos del corpus.

Gráfico 9. Fórmula de la inversa de la frecuencia de documentos.



Fuente: Elaboración propia en base a Rosati (2019).

Ambas métricas se agrupan en lo que se conoce como la matriz TF-IDF, donde cada término combina ambas métricas y por lo tanto nos hacemos de una medición tanto de importancia como de informatividad de un término a lo largo de un corpus de texto analizado.

Gráfico 10. Fórmula de la frecuencia de término - inversa de la frecuencia de documentos.

Fuente: Elaboración propia en base a Rosati (2019).

Al incorporar esta métrica a la matriz inicial, mejora los conteos crudos con los pesos de los TF_IDF a cada uno de los términos.

II.2 Modelización de tópicos

Como explican Silge y Robinson (2017) en los procesos de minería de datos orientados a texto, de manera frecuente existe una colección de documentos como artículos de noticias o entradas de blogs que es preciso dividir en grupos naturales para poder entenderlos de forma separada. La modelización de tópicos es una forma de clasificación no supervisada de documentos.

II.2.1 *Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) es un método particularmente popular de clasificación no supervisada para implementar modelización de tópicos. Trata a cada documento como una mixtura de tópicos, y cada tópico como una mixtura de palabras. Esto permite a los documentos "solaparse" uno con otro en términos del contenido, en lugar de estar separados en grupos discretos. El algoritmo *Latent Dirichlet Allocation* se guía por 4 supuestos clave:

1- Cada documento es una mixtura de tópicos

Se supone que cada documento contiene palabras que pertenecen a varios tópicos en proporciones particulares. Por ejemplo, en una modelización de dos tópicos, se podría decir que "el documento 1 contiene 90% del tópico A y 10% del tópico B, mientras que el documento 2 contiene 30% del tópico A y 70% del tópico B."

2- Cada tópico es una distribución de probabilidad sobre las palabras

Por ejemplo, al imaginar una modelización de dos tópicos sobre noticias argentinas donde existe un tópico "política" y otro "entretenimiento". Las palabras más probables en el tópico "política" pueden ser "presidente", "congreso", "gobierno", mientras en el tópico de "entretenimiento" las palabras más probables pueden ser "película", "televisión" y "actor". Es importante remarcar que las palabras pueden estar compartidas entre tópicos, por ejemplo, la palabra "presupuesto" puede aparecer en ambos tópicos.

3- Los tópicos preexisten a los documentos

Tópico es definido como una distribución de probabilidad sobre el vocabulario de todo el corpus. Todas las palabras van a tener una determinada probabilidad de pertenecer a un tópico y el modelo estima cuáles son aquellas palabras que tiene la probabilidad más alta de pertenecer a un tópico en cuestión. Sin embargo, será necesario definir a priori cuántos tópicos se desea encontrar en el corpus.

4- LDA asume un proceso generativo para poder realizar la estimación

El modelo va a intentar reconstruir el proceso de generación de los documentos del corpus. Para esto, el modelo asume que cada documento es producido por un proceso que puede resumirse de la siguiente forma:

1. se elige aleatoriamente una distribución a lo largo de la cantidad de tópicos que se ha definido como parámetro
2. luego para cada palabra del documento selecciona aleatoriamente un tópico de la distribución de tópicos del documento
3. selecciona aleatoriamente una palabra del tópico correspondiente (cada tópico era una distribución sobre palabras)

Básicamente esto es elegir un tópico aleatorio de la distribución de tópicos con X probabilidad y adentro de ese tópico sorteo palabras con la distribución de probabilidad de palabras específicas de ese tópico. Lo que se logra con esto es sortear las palabras más probables en los tópicos más probables.

El modelo probabilístico de tópicos estimado por *LDA* consiste en dos tablas (matrices). La primera tabla describe la probabilidad de seleccionar una palabra determinada al realizar un muestreo[16] de un tópico particular, la segunda tabla describe la chance de seleccionar un tópico particular cuando muestreamos un documento en particular (Lettier, 2018).

De forma más técnica, el algoritmo *LDA* asume que los documentos fueron generados de la siguiente manera:

1. Seleccionamos un conjunto de elementos (palabras).
2. Seleccionamos cuantos documentos (o 'compuestos') queremos.
3. Seleccionamos cuantos elementos (palabras) queremos por documento (una muestra de una distribución de Poisson).
4. Seleccionamos cuantos tópicos queremos.
5. Seleccionamos un número entre no-cero e infinito positivo y lo llamamos *alpha.*
6. Seleccionamos un número entre no-cero e infinito positivo y lo llamamos *beta.*
7. Construimos la tabla 'palabras-versus-tópicos'. Por cada columna, sacamos una muestra (giramos la rueda) de una distribución Dirichlet (que es una distribución de distribuciones) usando *beta* como valor de entrada. Cada muestreo completará cada columna de la tabla hasta sumar uno y entregará la probabilidad de cada palabra por tópico (columna).
8. Construimos la tabla 'documento-versus-tópicos'. Para cada fila, sacamos una muestra de una distribución Dirichlet usando *alpha* como valor de entrada. Cada muestra completará cada fila de la tabla, sumará uno, y entregará la probabilidad de cada tópico (columna) por documento.
9. Finalmente construimos el documento. Por cada documento: a) buscamos la fila que le corresponde en la tabla de 'documentos-versus-tópicos', b) muestreamos un tópico basado en la probabilidad de la fila, c) vamos a la tabla 'palabras-versus-tópicos', d) buscamos el tópico muestreado, e) muestreamos una palabra basado en la probabilidad

[16] Por muestreo referimos a la selección de un subconjunto de datos de una población estadística para estimar característica del conjunto total de la población a la que refiere.

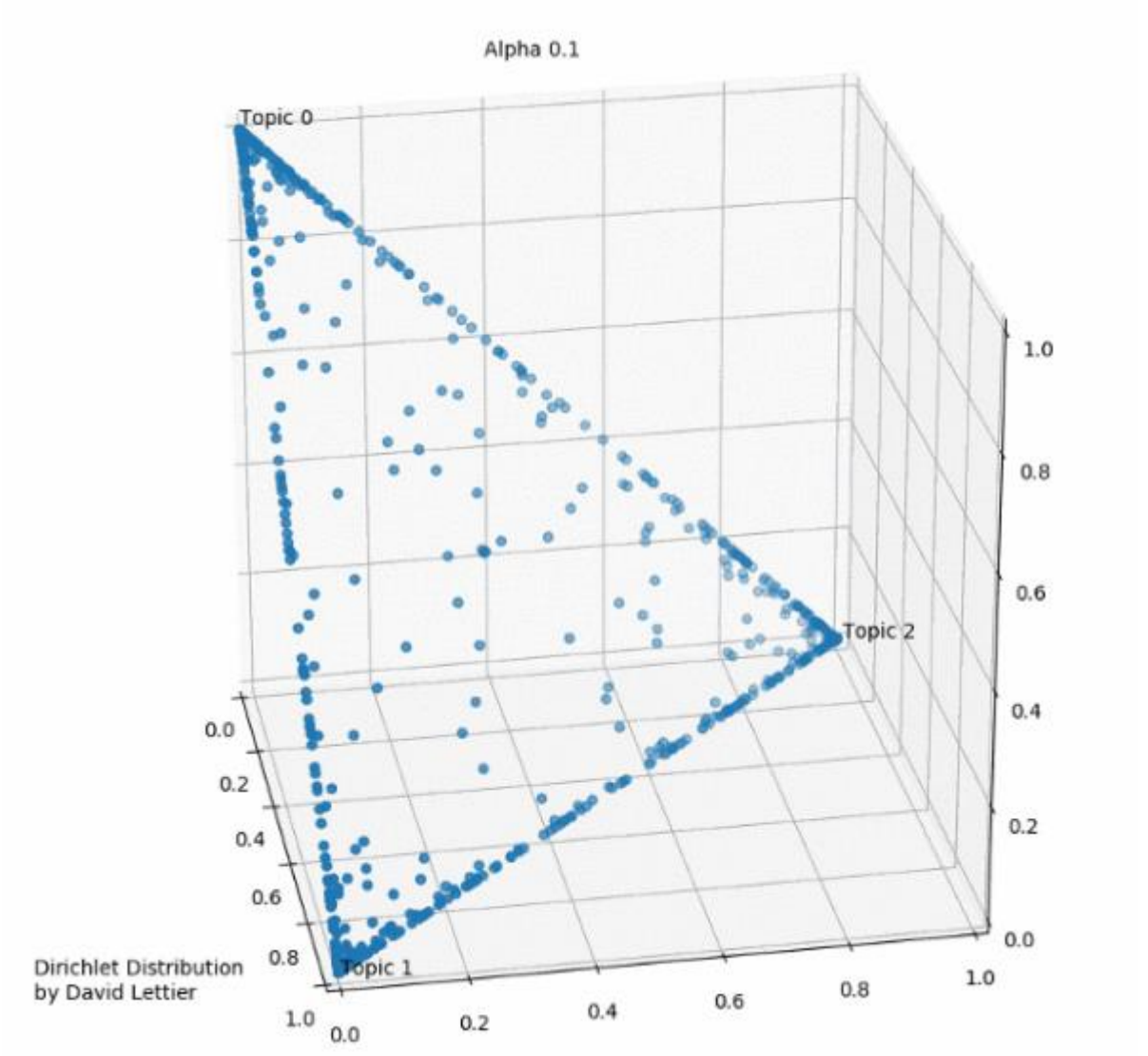
en la columna, f) repetimos el paso ‘b’ hasta que alcancemos la cantidad de palabras que queremos que ese documento contenga.

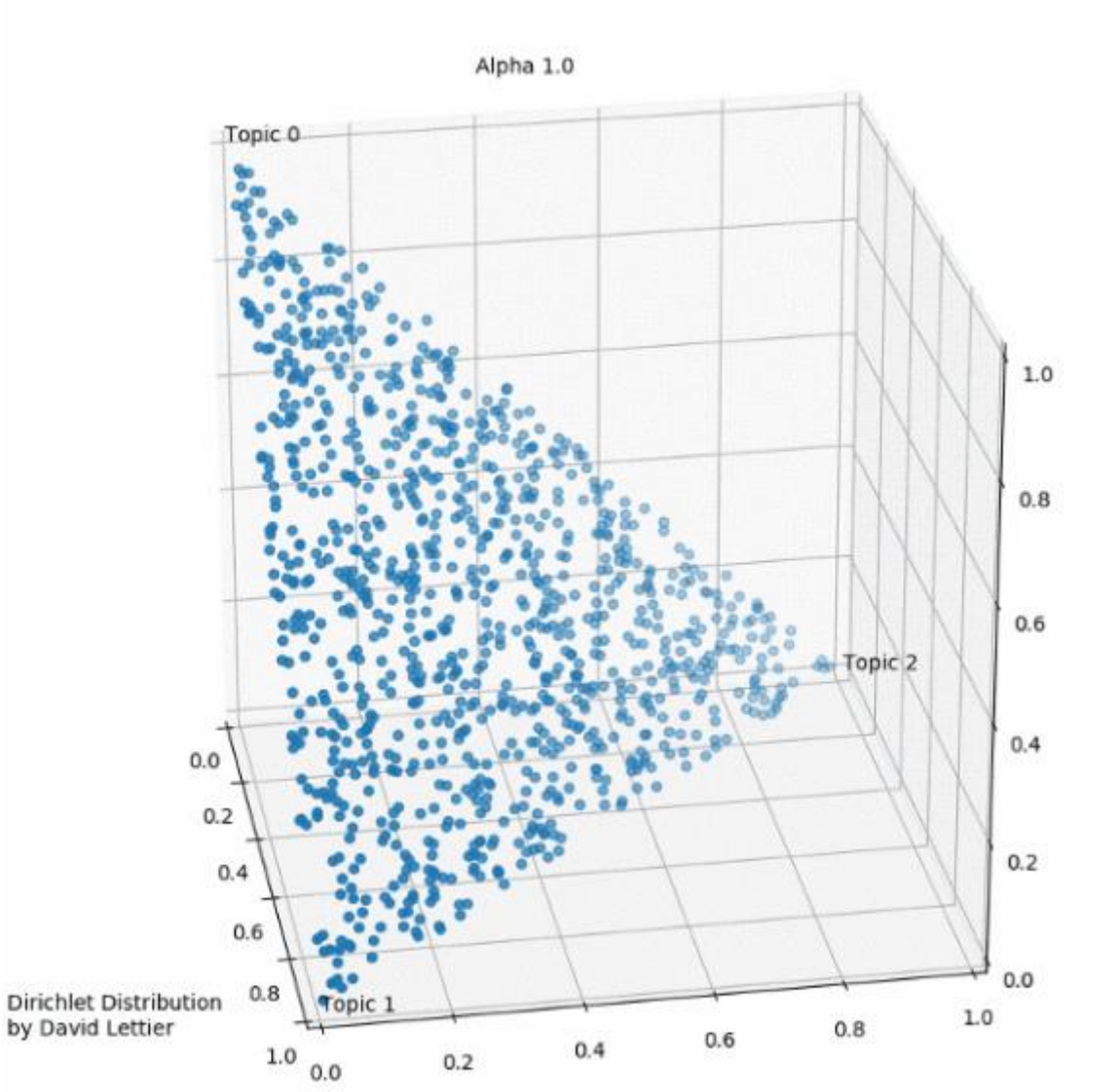
Este es entonces el modelo simplificado que *LDA* asume como proceso generativo de los documentos.

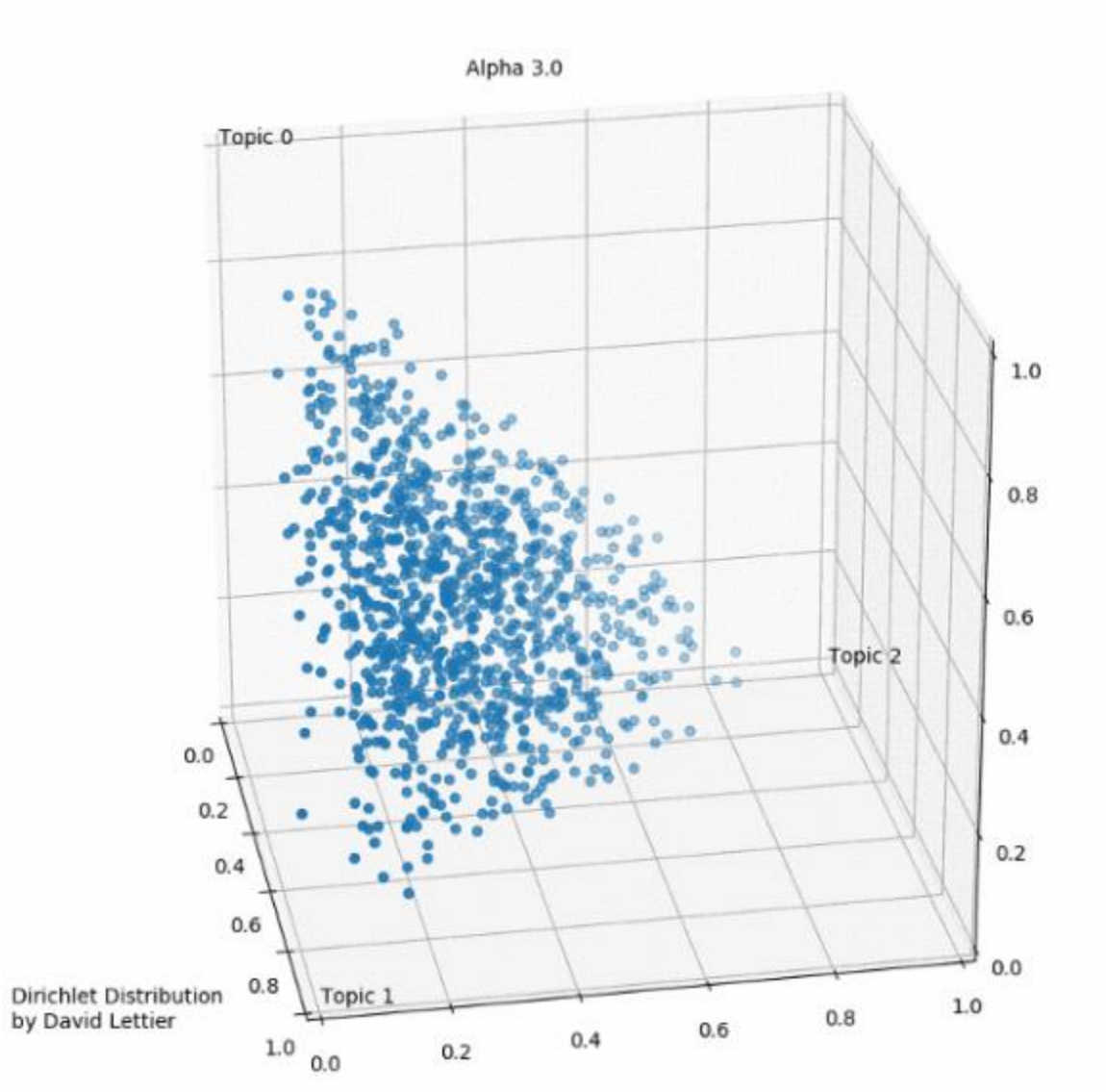
II.2.2 La distribución Dirichlet

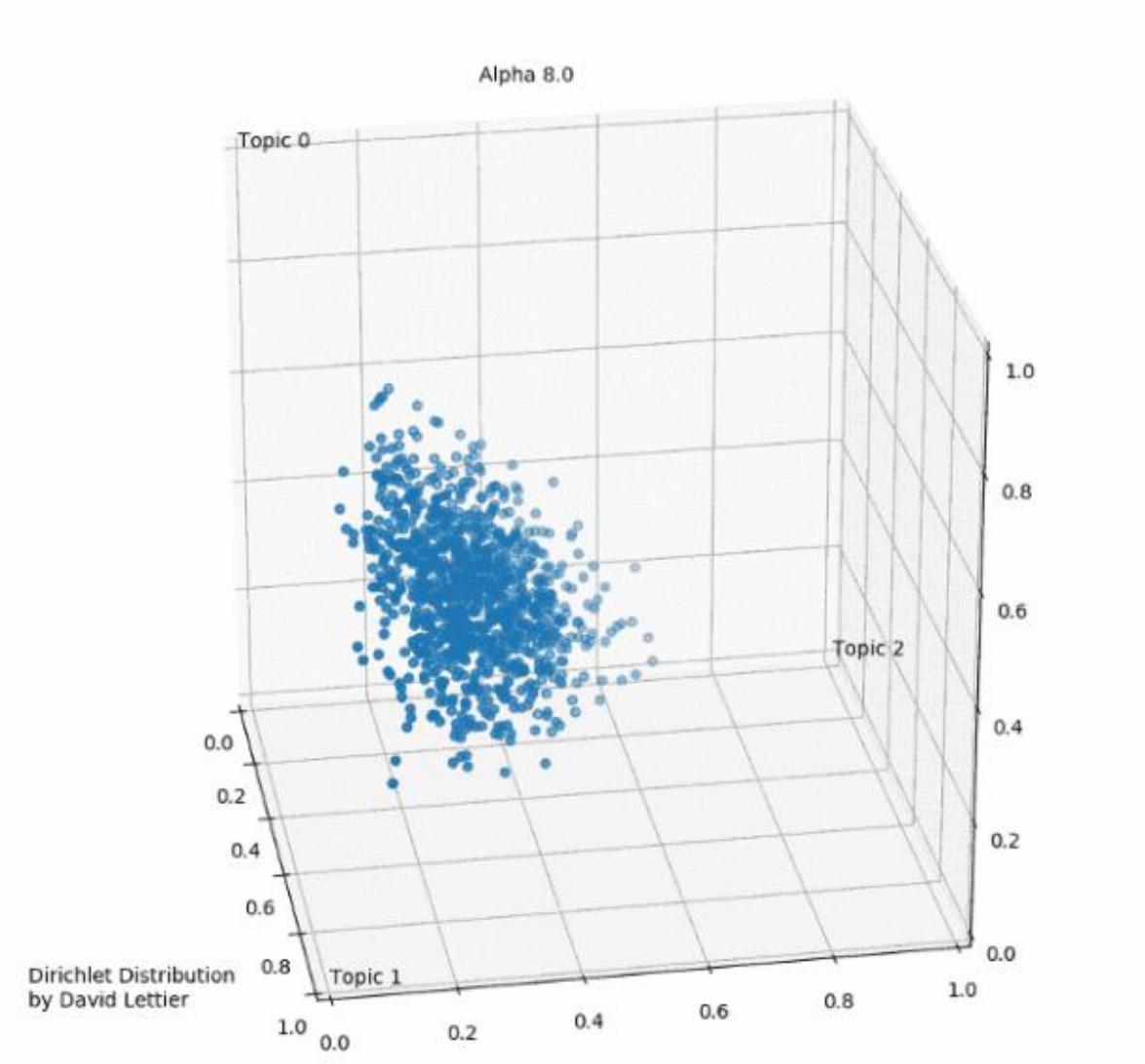
En las imágenes que se exponen a continuación se aprecia iteraciones tomando 1000 muestreos de una distribución Dirichlet usando un valor *alpha* que se va incrementando.

Gráfico 11. Secuencia de aumento del hiperparámetro *alpha*. Iteraciones tomando 1000 muestras de una distribución Dirichlet.




Alpha 1.0
Topic 0
Topic 1
Topic 2
0.0
0.2
0.4
0.6
0.8
1.0
Dirichlet Distribution
by David Lettier


Alpha 3.0
Topic 0
Topic 1
Topic 2
0.0
0.2
0.4
0.6
0.8
1.0
Dirichlet Distribution
by David Lettier



Fuente: Lettier (2018). *Your Guide to Latent Dirichlet Allocation.*

La distribución Dirichlet, nombre que toma en honor a su creador, el matemático alemán Johann Peter Gustav Lejeune Dirichlet, toma un número (llamado *alpha* en la mayoría de los casos) para cada tópico. En las imágenes expuestas (y para nuestro propósito didáctico) a cada tópico se le asigna el mismo valor *alpha* que se muestra en el título. Cada punto representa una distribución o mixtura de distribución de todos los tres tópicos como por ejemplo (1.0, 0.0, 0.0) o (0.4, 0.3, 0.3). Recordemos que cada muestreo tiene que sumar uno. Con bajos valores de *alpha* (menores a uno), la mayoría de los muestreos de distribución de tópico están en las esquinas (cerca de los tópicos). Para valores muy bajos de *alpha* es probable que obtengamos muestreos como (1.0, 0.0, 0.0), (0.0, 1.0, 0.0) o (0.0, 0.0, 1.0). Esto significa que un documento tendría solamente un tópico si estuviéramos construyendo un modelo probabilístico de tres

tópicos desde cero. Con un valor *alpha* igual a uno, cualquier espacio en la superficie del triángulo está uniformemente distribuido. Podríamos obtener de igual manera un muestreo favoreciendo solamente un tópico, un muestreo que sea una mixtura distribuida de manera equitativa de todos los tópicos, o algo en el medio de ambos casos. Para valores de *alpha* mayores a uno, el muestreo se empieza a congregar en el centro del triángulo. Esto significa que mientras *alpha* se incrementa, los muestreos serán probablemente más uniformes, esto quiere decir, representan una mixtura de tópicos equitativa de todos los tópicos. A los fines de la demostración se realizó una modelización de tres tópicos ya que es funcional en el contexto de una visualización de tres dimensiones, pero típicamente es mejor usar más tópicos al modelizar, dependiendo del corpus de texto que uno posea. Para pasar en limpio: *alpha* es el hiperparámetro que controla la mixtura de tópicos para cualquier documento dado, si reducimos el valor de *alpha* los documentos tendrán menos solapamientos de tópicos en el mismo documento, al aumentarlo, la mixtura de tópicos hacia adentro de un documento dado aumentará.

El hiperparámetro *beta* controla la distribución de palabras por tópico, al reducirla, los tópicos contendrán menos palabras, al aumentarla, los tópicos contendrán más palabras. Idealmente se busca que los documentos muestren una mixtura de algunos pocos tópicos por documento y las palabras que pertenezcan solamente a algunos tópicos.

Capítulo III: Implementación de la modelización de tópicos

III.1 Preparación de los datos

Scraping – en inglés literalmente "raspar" – hace referencia al proceso de extracción de datos de una fuente determinada. *Web scraping* es el proceso de extracción de datos de un sitio web. En principio, al tener definidos los sitios a analizar, fue necesario realizar una recolección de las noticias entre enero de 2015 y diciembre de 2016. Para esto, se requieren los enlaces o direcciones web que apunten a la producción de noticias en ese recorte temporal. El proyecto GDELT es esencial para realizar esta tarea. Tal como se desprende su página oficial[17], GDELT (*Global Database of Events, Language and Tone)* se presenta como un proyecto respaldado por Google Jigsaw[18] que monitorea la emisión de noticias web de casi todos los rincones del mundo en más de 100 idiomas. Al mismo tiempo, identifica personas, localidades, organizaciones, temas, fuentes, emociones, conteos, citas, imágenes y eventos, creando una plataforma abierta y gratuita para computar. Si bien son varios los servicios que GDELT ofrece, lo que interesa en este trabajo es la capacidad de indexar[19] la emisión de noticias web en territorios específicos. Por lo expuesto, es que se utiliza GDELT como la fuente de consulta al generar la lista de links del proceso electoral del 2015 en el recorte espacio temporal propuesto.

La base de datos GDELT es accesible a través de la plataforma de computación en la nube de *Google: Google Cloud Platform (GCP).* De los muchos servicios que ofrece tal plataforma, interesa particularmente el servicio *Google Big Query*, un almacén de datos multinube de alta escalabilidad, sin servidor (*serverless)* desarrollado para analizar grandes volúmenes de datos a altas velocidades y sin sobrecarga operativa mediante ANSI SQL[20].

Dicha consulta devuelve 502.916 noticias entre las 00:00hs del 01/01/2015 a las 00:00hs del 01/01/2017 para los medios *Télam, La Nación, Clarín, Perfil, Infobae, MinutoUno y Página 12.* La consulta procesó casi 700GB de información y demoró 12.4 segundos.

[17] https://www.gdeltproject.org/

[18] Google Jigsaw forma parte de Alphabet, el grupo propietario de Google. Según su sitio oficial se trata de una incubadora de tecnología con el objetivo de abordar algunos de los retos más difíciles relacionados con la seguridad internacional en el mundo actual, cómo luchar contra la censura en Internet, reducir las amenazas de ataques digitales, contrarrestar la violencia del extremismo o proteger del ciberacoso a las personas. Para conocer más: https://jigsaw.google.com/

[19] Insertar en una base de datos artículos de noticias al momento de su publicación, en conjunto otra información relevante.

[20] https://cloud.google.com/bigquery

Posteriormente a la obtención de los enlaces, se desarrolló un *web scraper* en lenguaje Python utilizando las librerías *dataset*[21] para la conexión a una base de datos *Sqlite* y *beautiful soup*[22] y *newspaper*[23] para la extracción del texto. El script se implementó sobre un servidor del servicio de computación en la nube de *Amazon (Amazon Web Services)*, particularmente utilizando el servicio *Amazon Lightsail*, que provee de equipos virtuales para el alojamiento de diferentes servicios.

Finalizado el procesamiento del script para la extracción de texto, se obtiene el primer dato concreto del análisis. La base de datos final consta de 466.754 observaciones o filas. El motivo de esas 36.162 noticias faltantes está vinculado a la indisponibilidad del enlace en internet, en otras palabras, los enlaces o están caídos (los sitios no responden al requerimiento) o las noticias no presentan texto, o lo presentan con errores. El script fue explícitamente programado para eliminar de la base de datos los sitios que presenten tales errores[24].

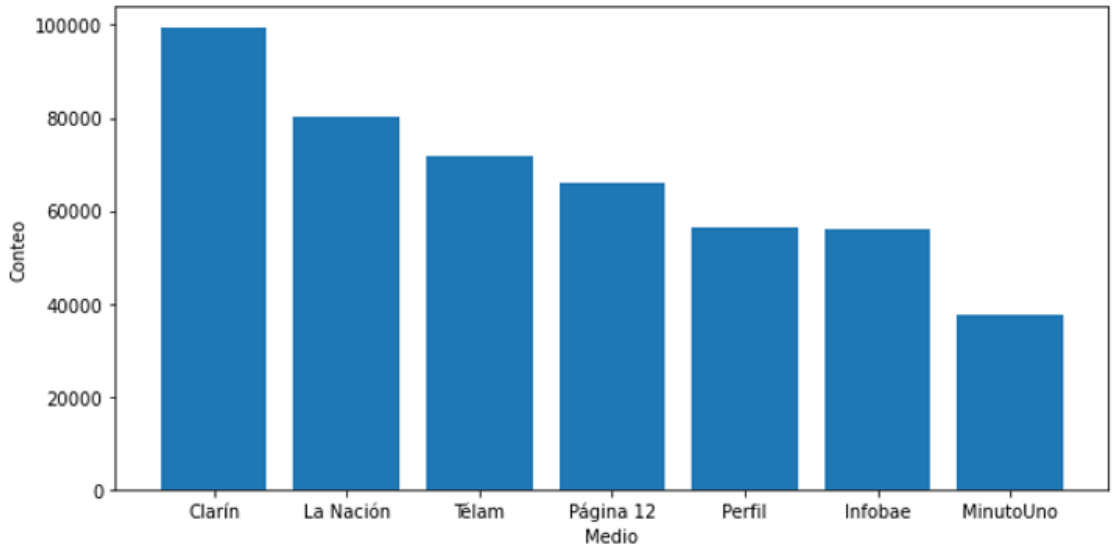
Para el posterior análisis de la base de datos se utilizará una librería de amplia difusión en el campo computacional: *pandas*[25]. Está específicamente desarrollada para ayudarnos a explorar, limpiar y procesar datos tabulares.

Como se puede observar en el cuadro 1 que muestra las primeras 5 entradas de la base de datos final, existen 4 columnas (título, texto, fecha, link). La columna *"*fecha*"* aloja la información de la fecha y hora de publicación de la noticia, mientras que *"*link*"* es la dirección del enlace web.

[21] https://dataset.readthedocs.io/
[22] https://www.crummy.com/software/BeautifulSoup/bs4
[23] https://newspaper.readthedocs.io/en/latest/
[24] Consultar diagrama de flujo en Anexo III
[25] https://pandas.pydata.org/

III.2 Análisis exploratorio

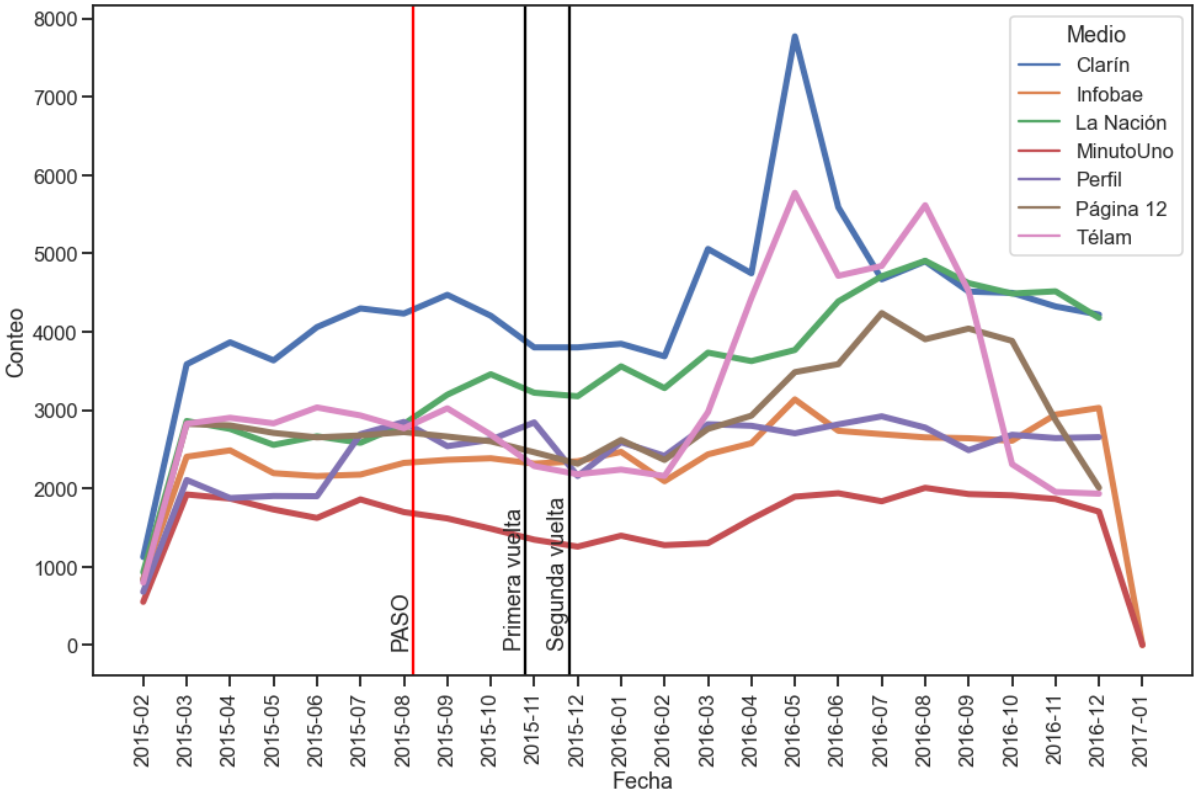
Gráfico 12. Cantidad de notas por medio de comunicación. De enero de 2015 a diciembre de 2016.



Fuente: Elaboración propia en base a GDELT.

En el gráfico 12 se observa la disparidad en la producción periodística de los medios estudiados, destacando la producción de *Clarín y La Nación*, seguido por *Télam, Pagina 12, Perfil, Infobae y MinutoUno*.

Gráfico 13. Cantidad de notas por medio de comunicación. De enero de 2015 a diciembre de 2016.



Fuente: Elaboración propia en base a GDELT.

El gráfico 13 muestra como *Clarín* presenta un pico de producción periodística entre el 05/2016 y el 06/2016. *Infobae* exhibe una relativa estabilidad en los números de producción periodística. *La Nación* presenta un alza en su producción entre el 08/2016 y 07/2016. *MinutoUno* presenta una leve alza el mes 07/2015 y otro entre el 05/2016 y 06/2016. Perfil también presenta una relativa estabilidad en su producción, salvo por una leve baja en su producción entre el mes 04/2015 y 06/2015. *Página 12* muestra un pico de producción entre los meses 07/2016 al 10/2016. Por último, Télam, presenta un pico bastante claro entre 05/2016 y el 08/2016.

III.3 Modelización y visualizaciones

Para la implementación de la modelización se optó por utilizar el paquete *scikit-learn*[26] para Python. Se intentó buscar un valor para *k* (número de tópicos) que devuelva un conjunto de tópicos interpretables. El anexo II muestra la lista de tópicos detectados para *k=10* como así también la nominación escogida para cada uno de esos tópicos.

[26] https://scikit-learn.org/

Es importante recordar que los tópicos se definen como una distribución de probabilidad sobre las palabras del vocabulario, una misma palabra tiene una cierta probabilidad de pertenencia a todos los tópicos, las diferencias entonces son de carácter relativo. Ciertas palabras pertenecen a ciertos tópicos con mayor probabilidad que otras (Rosati, 2020). Esto permite que exista un solapamiento entre términos a lo largo de los tópicos. Cada tópico que se presenta en el anexo muestra las 30 palabras con mayor probabilidad de pertenencia a cada tópico.

Se constata que el tópico 1 posee términos que refieren al arte, cine y espectáculo. El tópico 2 y el 6 hablan ambos de hechos violentos, pero el 6 reúne los vinculados a inseguridad, mientras que el 2 muestra noticias trágicas o de relación con hechos violentos impactantes. El tópico número 3 hace referencia a noticias sobre política exterior o política internacional. El tópico 4 reúne noticias que tratan cuestiones sobre personalidades públicas vinculadas a los espectáculos o medios de comunicación. El tópico 5 reúne hechos de corrupción y noticias vinculados al ámbito del sistema de justicia, tanto avance de causas, como denuncias o declaraciones que refieren a causas que tramitan en la justicia. El tópico 7 reúne noticias vinculadas a los deportes en general. El tópico 8 tiene un carácter residual (pocas noticias presentan media o alta probabilidad de pertenencia este tópico), si bien de los términos puede intuirse la existencia de noticias vinculadas a la justicia, el análisis de las noticias que componen el tópico es poco claro, reúne tanto noticias de deportes, como hechos violentos o columnas de opinión, por lo que se optó por nominarlo como ‘no interpretable’ y no considerarlo en el análisis. El tópico 9 reúne noticias de carácter electoral, vinculados de manera directa o indirecta a cuestiones proselitistas. Finalmente, el tópico 10 reúne las noticias de tipo económicas. El tópico 2 y el tópico 6 han sido reunidos en uno solo, nominado “Hechos violentos/Inseguridad”. Por otro lado, el tópico no interpretable no forma parte de las visualizaciones por no aportar información valiosa al análisis.

Cuadro 4. Tabla de tópicos detectados.

Tópico	Denominación
Tópico 1	*Cine, arte y espectáculos*
Tópico 2	*Política exterior*
Tópico 3	*Farándula*
Tópico 4	*Causas judiciales y hechos de corrupción*
Tópico 5	*Deportes*
Tópico 6	*Elecciones*
Tópico 7	*Economía*
Tópico 8	*Hechos violentos/Inseguridad*

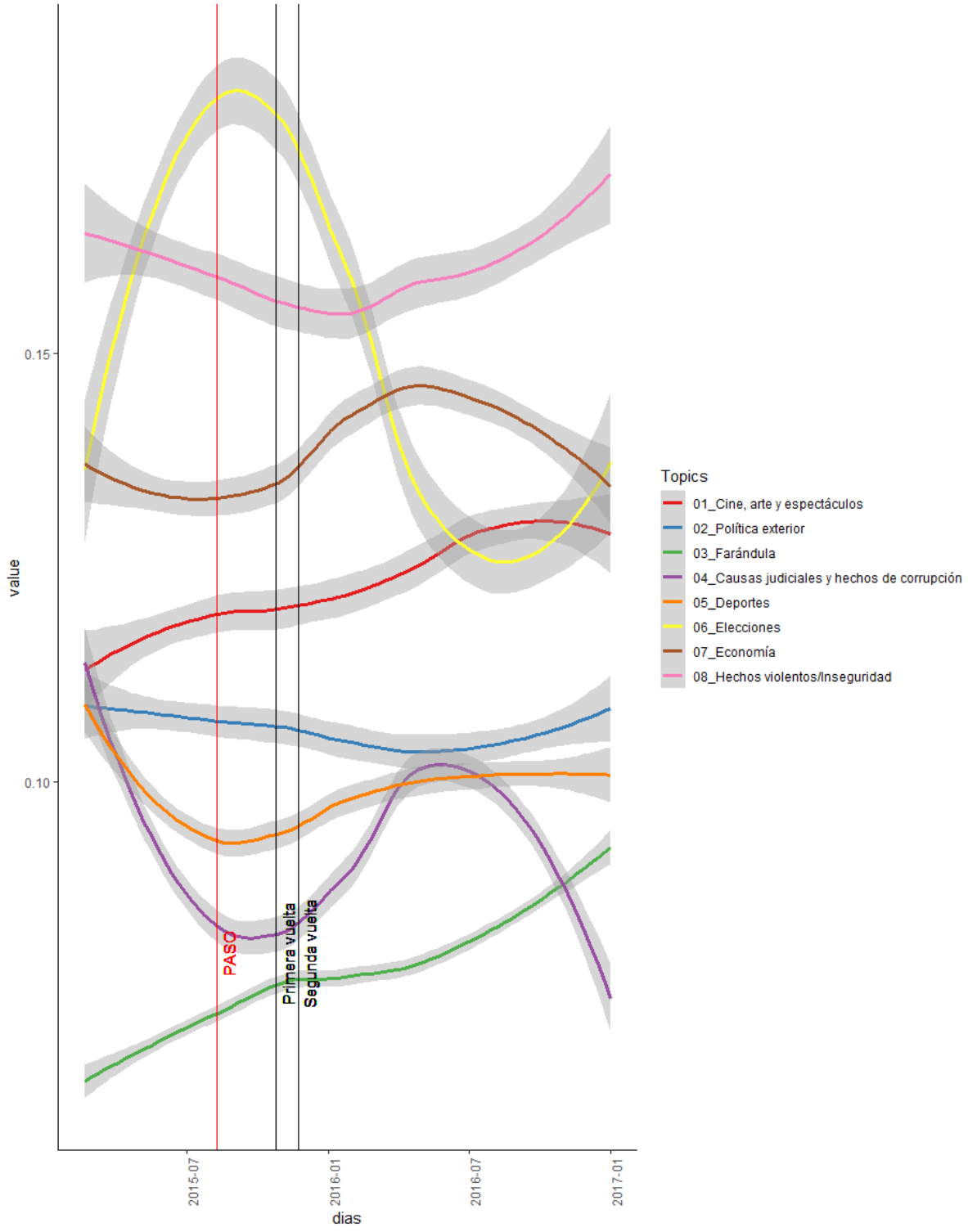
Fuente: Elaboración propia.

La visualización que se expone en el gráfico 14 reúne a todos los medios de comunicación en un mismo gráfico y calcula la media de cada tópico a lo largo de todas las noticias de cada día. Dado que se tiende a observar muchas oscilaciones se aplicó un suavizado de las series temporales de cada tópico usando el método GAM[27].

Se corrobora cómo el tópico vinculado a noticias sobre elecciones y política electoralistas en general muestra un pico en su cobertura que es compatible con el acto electoral en sí mismo. El valor máximo del tópico viene luego de pasados unos días posterior a las elecciones Primarias Abiertas Simultáneas y Obligatorias (PASO), para luego decrecer pasadas las elecciones generales. Es posible corroborar un cambio en la composición de los tópicos más relevantes en los medios estudiados dentro del recorte espacio temporal planteado con anclaje en el acto electoral inaugural del período 2015, las PASO.

[27] Por sus siglas en inglés Generalized Additive Models. Se trata de un modelo lineal generalizado en el que la variable respuesta depende linealmente de una serie de funciones de suavizado desconocidas sobre las variables independientes. El objetivo es estimar estas funciones (Hastie y Tibshiani 1990).

Gráfico 14. Evolución de los tópicos. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.



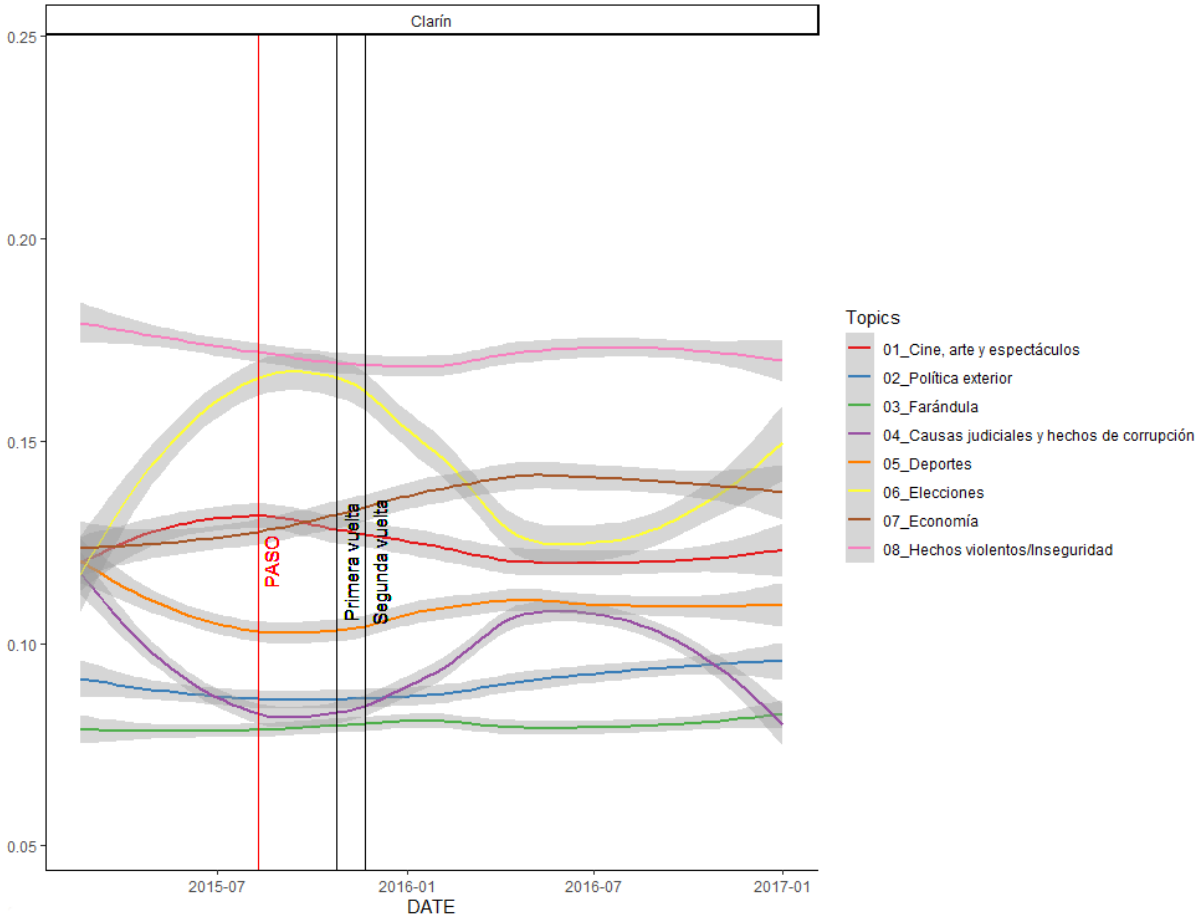
Fuente: Elaboración propia.

Para resolver el objetivo específico planteado se realizará una desagregación de la evolución de la composición de la media de los tópicos detectados por medio de comunicación. La

desagregación por medio de comunicación apunta a evaluar si existe o no diferencias en la presencia de los tópicos detectados por sobre el conjunto de las noticias relevadas cuando se analiza por medio de comunicación.

Clarín

Gráfico 15. Evolución de los tópicos en Clarín. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.



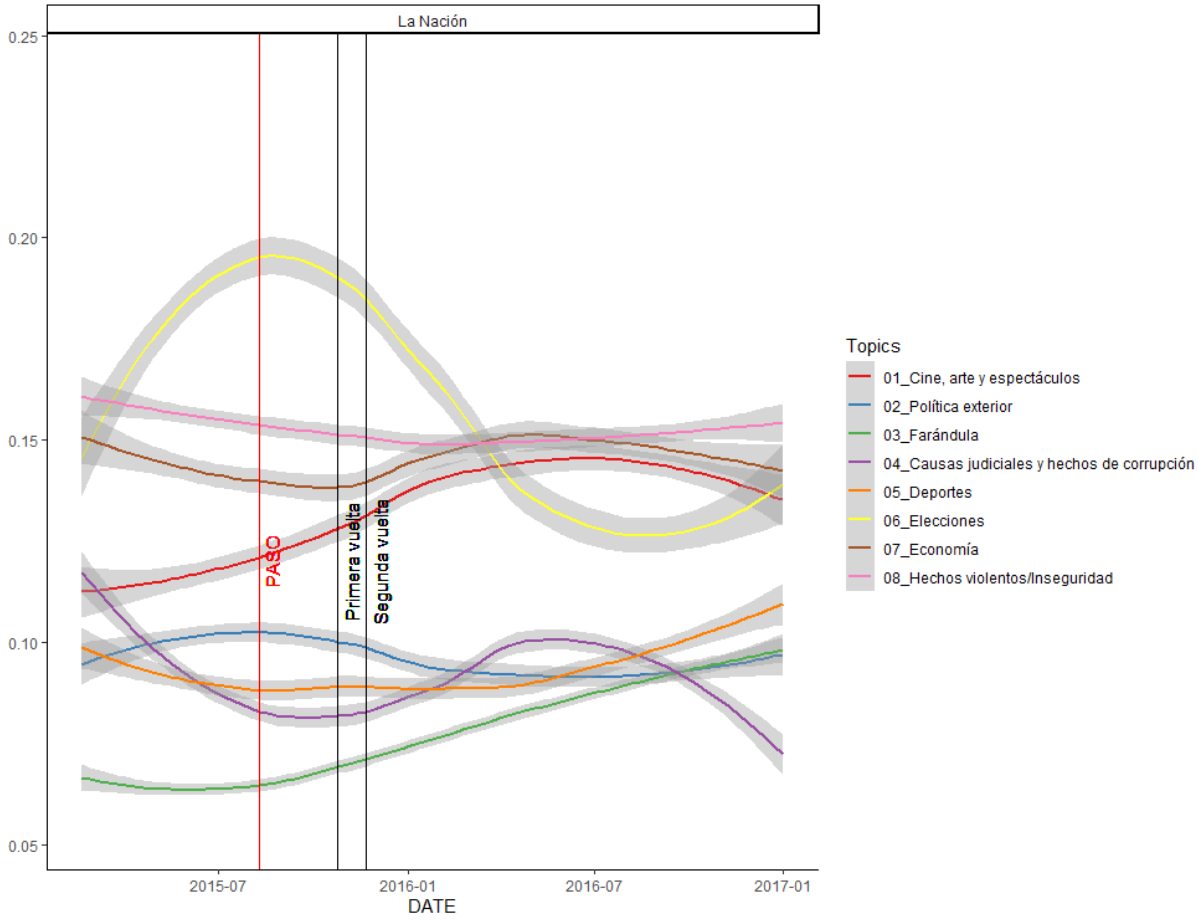
Fuente: Elaboración propia.

La evolución de la media de la composición de tópicos del diario Clarín presenta características que no se verifican en otros medios de comunicación. Los tópicos de violencia e inseguridad tienen una fuerte presencia a lo largo de todo el periodo estudiado. La evolución del tópico sobre elecciones se inscribe dentro de la tendencia general sin desagregación, exhibe un pico entre las PASO y las elecciones generales, para luego perder presencia, recobrándola sobre el

final del período. Las noticias vinculadas a la economía registran un alza posterior a las elecciones generales, convirtiéndose en tópico predominante por un breve período de tiempo entre 05/2016 y el 10/2016. Por último, es posible señalar un alza sostenida del tópico vinculado a causas judiciales y hechos de corrupción posterior a las elecciones de 2015, para luego perder presencia a final del período.

La Nación

Gráfico 16. Evolución de los tópicos en La Nación. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.

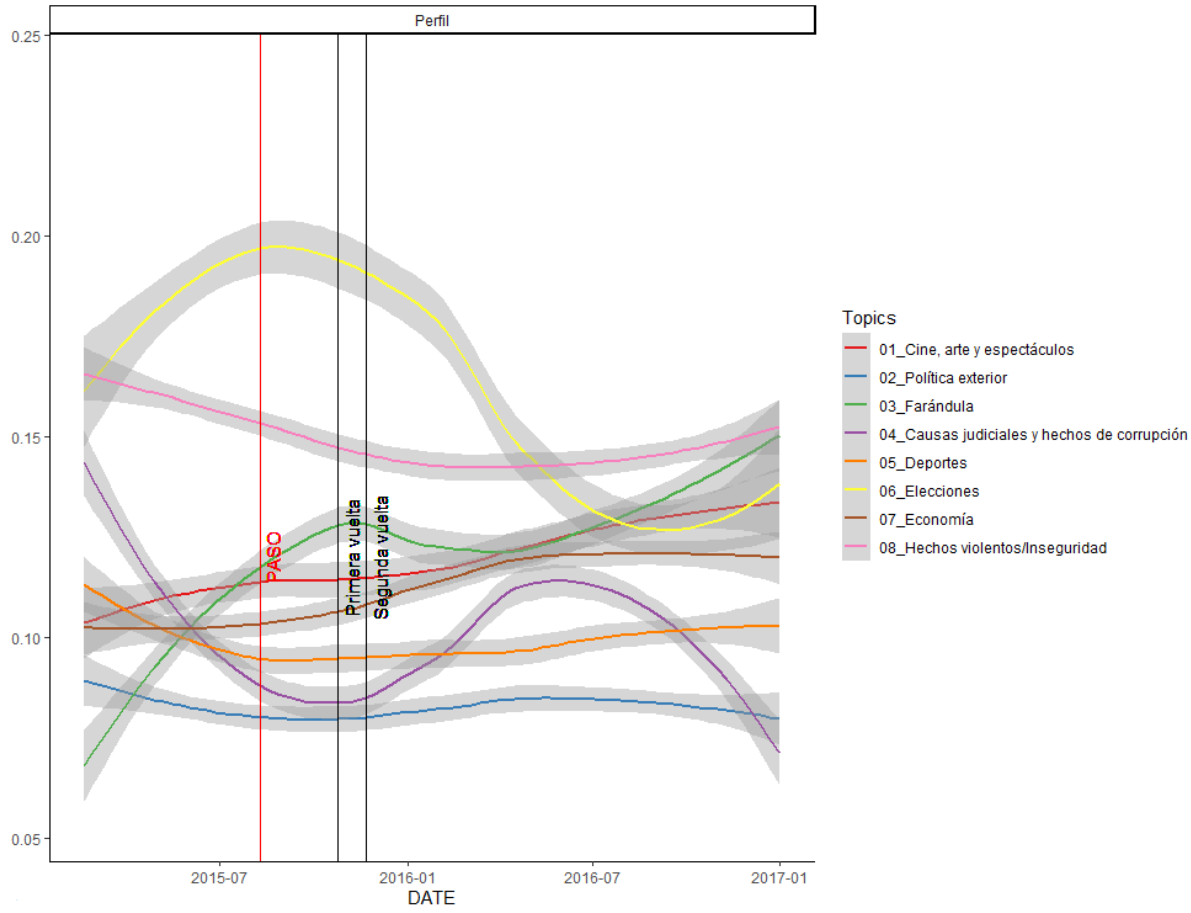


Fuente: Elaboración propia.

La Nación presenta un marcado predominio del tópico sobre elecciones o política electoral y también se inscribe en la tendencia general, presenta un pico durante las PASO y a posterior el tópico pierde presencia para ser superado por los tópicos económicos y violencia.

Perfil

Gráfico 17. Evolución de los tópicos en Perfil. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.

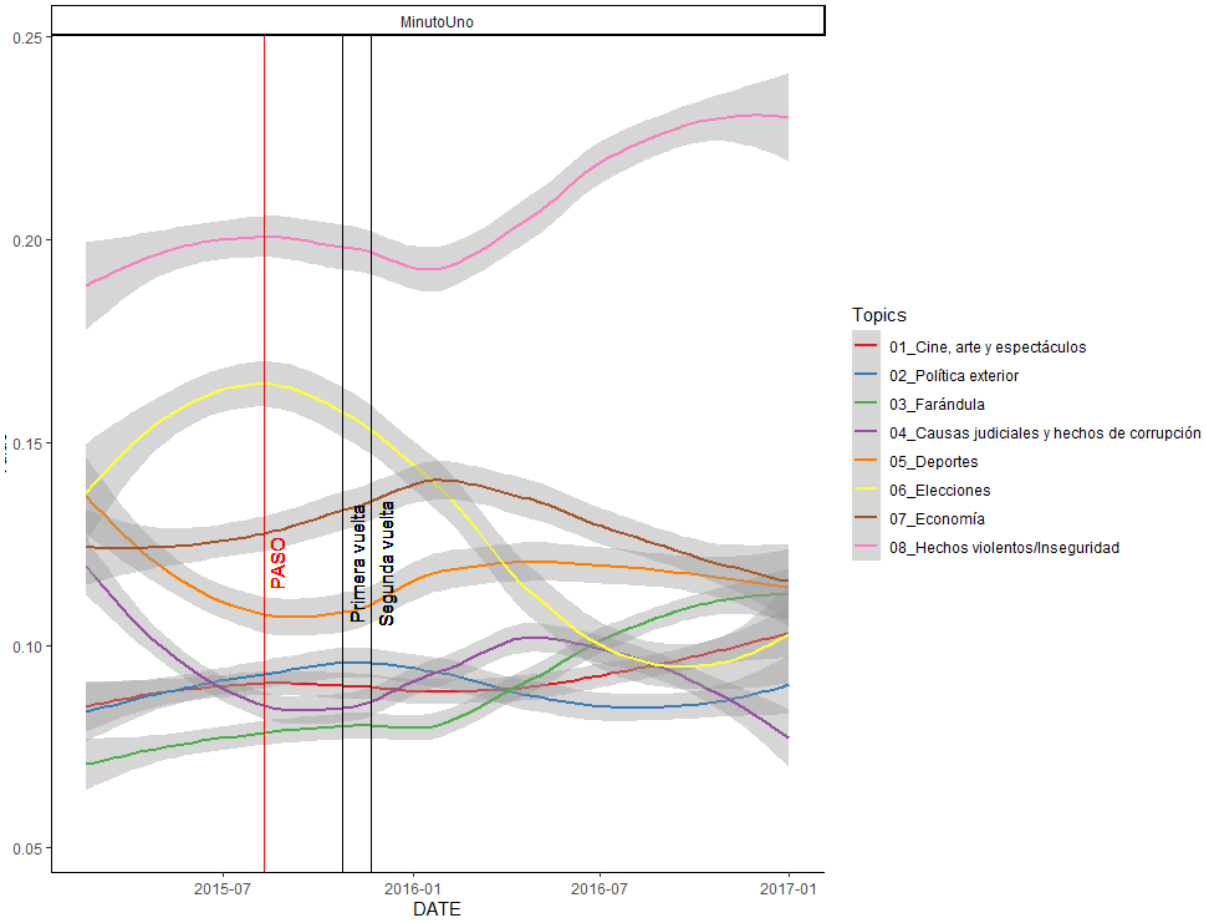


Fuente: Elaboración propia.

Como dato destacable de los tópicos que muestra el grupo perfil, se verifica una presencia fuerte respecto al tópico sobre farándula, es debido a que uno de los subdominios de Grupo Perfil es el portal Caras. En su agrupación por medio, el tópico farándula tiene una fuerte presencia, ya que Caras se encuentra en el dominio 'caras.perfil.com.ar'. Dejando de lado el tópico electoral, preponderan los tópicos de violencia, cine, arte y espectáculos y economía.

MinutoUno

Gráfico 18. Evolución de los tópicos en Minuto Uno. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.

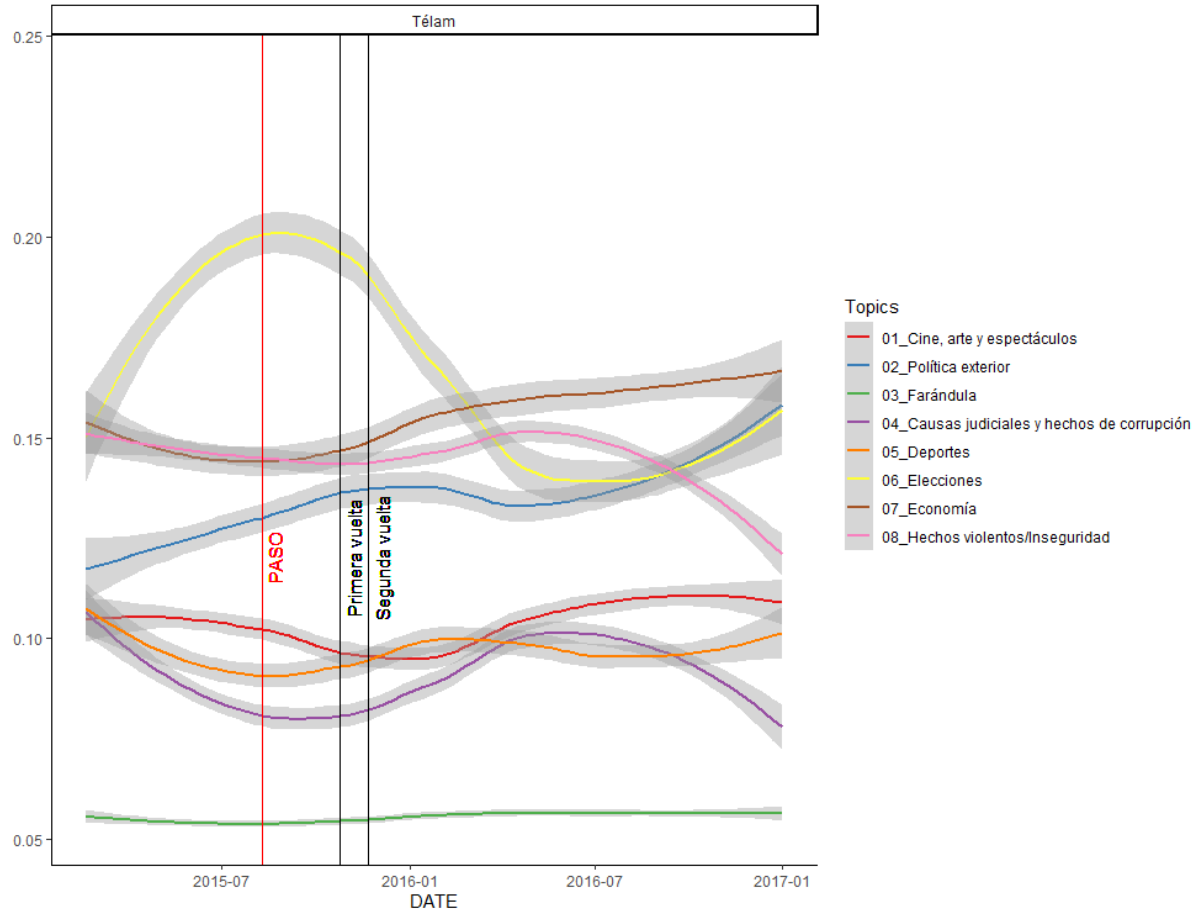


Fuente: Elaboración propia.

Se destaca la presencia del tópico sobre hechos violentos e inseguridad, el más fuerte y marcado de todos los medios. Destacan luego los tópicos electorales, con su respectivo decrecimiento luego del acto electoral que se verifica de manera transversal en todos los medios. Predominan también los tópicos económicos y deportes.

Télam

Gráfico 19. Evolución de los tópicos en Télam. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.

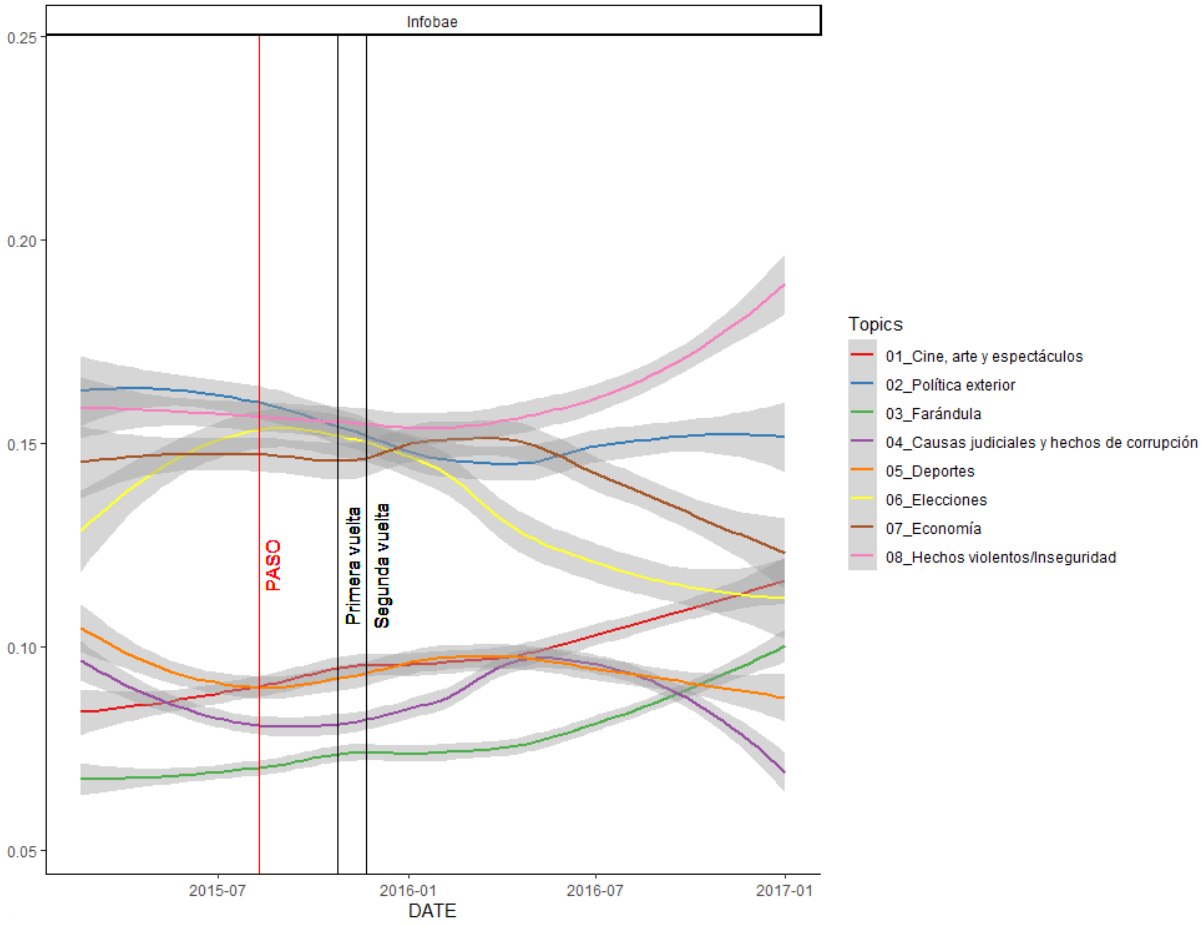


Fuente: Elaboración propia.

Se verifica la presencia dominante del tópico electoral en contexto de las elecciones PASO y las elecciones generales, para decrecer posteriormente siguiendo la tendencia general de evolución de tópicos. Dominan la composición de la media de tópicos los vinculados a economía y a hechos violentos, para luego mostrar una presencia fuerte sobre noticias de política exterior.

Infobae

Gráfico 20. Evolución de los tópicos en Infobae. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.

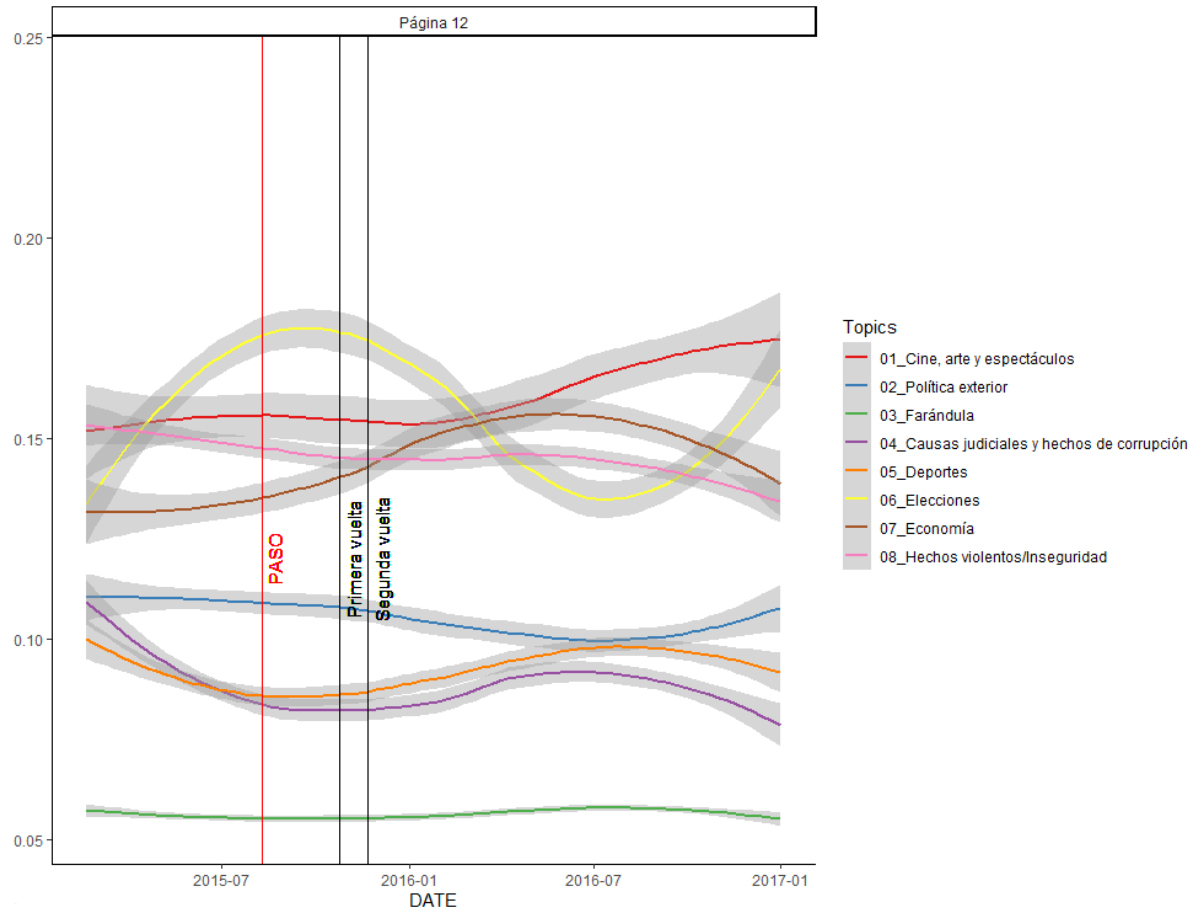


Fuente: Elaboración propia.

Infobae muestra una de las evoluciones de tópicos que menos se ajustan a las tendencias generales. Tal como Clarín, predominan las noticias vinculadas a hechos violentos o inseguridad, para luego mostrar una fuerte presencia de noticias de carácter político internacional, tendencia solo compartida con Télam. Luego, presenta un fuerte predomino del tópico vinculado a la economía.

Página 12

Gráfico 21. Evolución de los tópicos en Página 12. Media diaria (suavizada por GAM) de la composición de los tópicos de noticias.



Fuente: Elaboración propia.

Finalmente, Página 12 muestra característica propias que no es posible verificar en otros medios. El tópico sobre cine, arte y espectáculos tiene una presencia dominante. Por otra parte, los tópicos sobre economía y hechos violentos o de inseguridad dominan la evolución de tópicos.

Tomando en consideración la totalidad del período estudiado, los tres tópicos que lideran la cobertura mediática en la desagregación por medio son el tópico 8, sobre hechos violentos e inseguridad, seguido del tópico 7, economía y finalmente el tópico 6, elecciones. Resulta interesante advertir las trayectorias del tópico sobre elecciones y el tópico sobre causas

judiciales y corrupción, en general, el aumento de un tópico corresponde con el descenso del otro, cuando el tópico electoral desciende, el tópico sobre causas judiciales y corrupción compensa la pérdida de cobertura. Esto sucede en todos los casos estudiados, con mayor o menor grado de correlación.

Conclusiones

En la introducción del presente trabajo se expusieron algunas problemáticas comunes al estudiar, desde las ciencias sociales, fuentes de bajo o medio grado de estructuración como diarios, revistas, entrevistas y diversos tipos de documentos escritos. Además se señaló el problema de la complejidad de replicación de los resultados y los problemas en término de la escalabilidad del análisis. Se mencionó las posibilidades que otorga el uso de herramientas computacionales (*web scraping y Natural Language Processing*) a los fines de atender a las dificultades previamente planteadas.

Inicialmente se propuso analizar la agenda mediática en el contexto de las elecciones de Argentina en 2015, priorizando la implementación de técnicas computacionales para tal fin. De hecho, una de las preguntas que fueron el origen de la presente tesis (la tercera) se refiere a la viabilidad de las técnicas de *web scraping* y *NLP* para abordar estas temáticas. En ese sentido, fue posible armar un corpus de texto de más de 400.000 mil noticias recolectadas entre enero de 2015 y diciembre de 2016. Tal corpus empleó un grado de automatización en su proceso relativamente elevado para los estándares de la disciplina y del campo de estudios en particular:

- se realizó una consulta a la base de datos GDELT
- se *scrapearon* las noticias mediante scripts programados en el lenguaje Python
- se procesó la información obtenida utilizando el mismo lenguaje para su procesamiento y análisis posterior
- se realizó un modelado de tópicos utilizando el modelo generativo *Latent Dirichlet Allocation* obteniendo ocho tópicos predominantes en la agenda de los medios de comunicación estudiados

Se encuentra disponible el código e indicaciones de implementación en **https://tomasebm.github.io/topicmodeling/**. Es importante destacar la posibilidad de replicación, los scripts publicados son los mismos que fueron utilizados para generar los resultados del presente trabajo, como así también la implementación indicada. Asimismo, se encuentra a disposición el set de datos utilizado en el repositorio mencionado.

Previamente se había definido el concepto de tópico como una serie de acontecimientos relacionados con el tratamiento periodístico que se agrupan en una categoría más amplia. Estos acontecimientos, directamente observables en la superficie del discurso, constituyen los tópicos, es decir, etiquetas que resumen el dominio de las experiencias sociales incluidas en el relato (Pan y Kosicki, 1993). Se ha propuesto como una razonable aproximación a una operacionalización técnica del concepto abstracto el entendimiento de tópico que se desprende de la implementación de una modelización de tópico (*Latent Dirichlet Allocation* en este caso), proveniente del campo del aprendizaje automático.

En relación a la primera pregunta que orientó esta tesis, se logró individualizar tópicos vinculados a cine, arte y espectáculos, política exterior, farándula, causas judiciales y hechos de corrupción, deportes, elecciones, economía y hechos violentos e inseguridad. A su vez, la detección de tópicos que realiza el modelo LDA permite el solapamiento de tópicos, lo que habilita la posibilidad de evaluar de qué tópicos y en qué medida están incluidos una determinada cantidad de tópicos en un documento dado.

A su vez, se realizó una aproximación a la segunda pregunta (sobre la evolución temporal de los temas). Las visualizaciones confeccionadas en base al promedio de la evolución de la media de tópicos para el corpus objeto de análisis muestra cómo el tópico número 6, vinculado a noticias proselitistas o política electoral, ha dominado la agenda en los tres meses anteriores a las Primarias Abiertas Simultáneas y Obligatorias (PASO) siendo el segundo tópico predominante entre enero y junio de 2015 solo superado por el tópico de hechos violentos e inseguridad. El pico de cobertura se da precisamente transcurrida las PASO de 2015 y, una vez realizada la primera y segunda vuelta de la elección general, el tópico deja de dominar la agenda de los medios. Como consecuencia de esta caída la agenda queda principalmente dominada por los tópicos número 8 (hechos violentos e inseguridad) en primer lugar y luego por el número 7 (economía). El análisis desagregado por medio de comunicación aportó información adicional sobre características propias de las agendas de los medios estudiados. La alta presencia del tópico sobre hechos violentos e inseguridad en *Clarín*, como también la alta presencia del tópico de política exterior en *Infobae* o el tópico farándula en *Perfil* son algunas de las observaciones que se desprenden de las visualizaciones confeccionadas.

Existieron dimensiones que han quedado excluidas del análisis pero son importantes mencionar. No se ha realizado ningún análisis respecto a la disposición del contenido en las páginas de inicio o *homepage* de los sitios estudiados, tal como lo explican Zunino y Grilli Fox

(2019) la ubicación es un criterio clásico de jerarquía informativa que los medios digitales utilizan. Capturar esa jerarquía es importante en función de definir la cobertura sobre un tópico determinado. Existen herramientas con las que es posible analizar la disposición de un sitio web, tanto análisis de imágenes como análisis de estructuras HTML. Otro factor importante omitido que se encuentra en Koziner (2020) es la incorporación del tráfico de usuarios sobre el sitio web en diferentes rangos horarios, lo cual es otro elemento importante que nos puede arrojar información valiosa respecto a la cobertura que reciben determinados tópicos a lo largo de los días.

Finalmente, se propone a futuro indagar sobre la posibilidad de complementariedad entre metodologías tradicionales y metodologías computacionales aplicando técnicas y métodos del aprendizaje automático para mejorar los resultados en los análisis de fuentes tales como diarios, revistas, sitios web o diversos documentos. Queda también el desafío de estudiar diferentes maneras de optimizar el proceso en función de obtener los mejores resultados. Desde el proceso de ingeniería de datos hasta la selección de hiperparámetros para implementar la modelización de tópicos, como así también otros algoritmos de modelización de tópicos y visualizaciones que puedan sintetizar de mejor manera los resultados obtenidos.

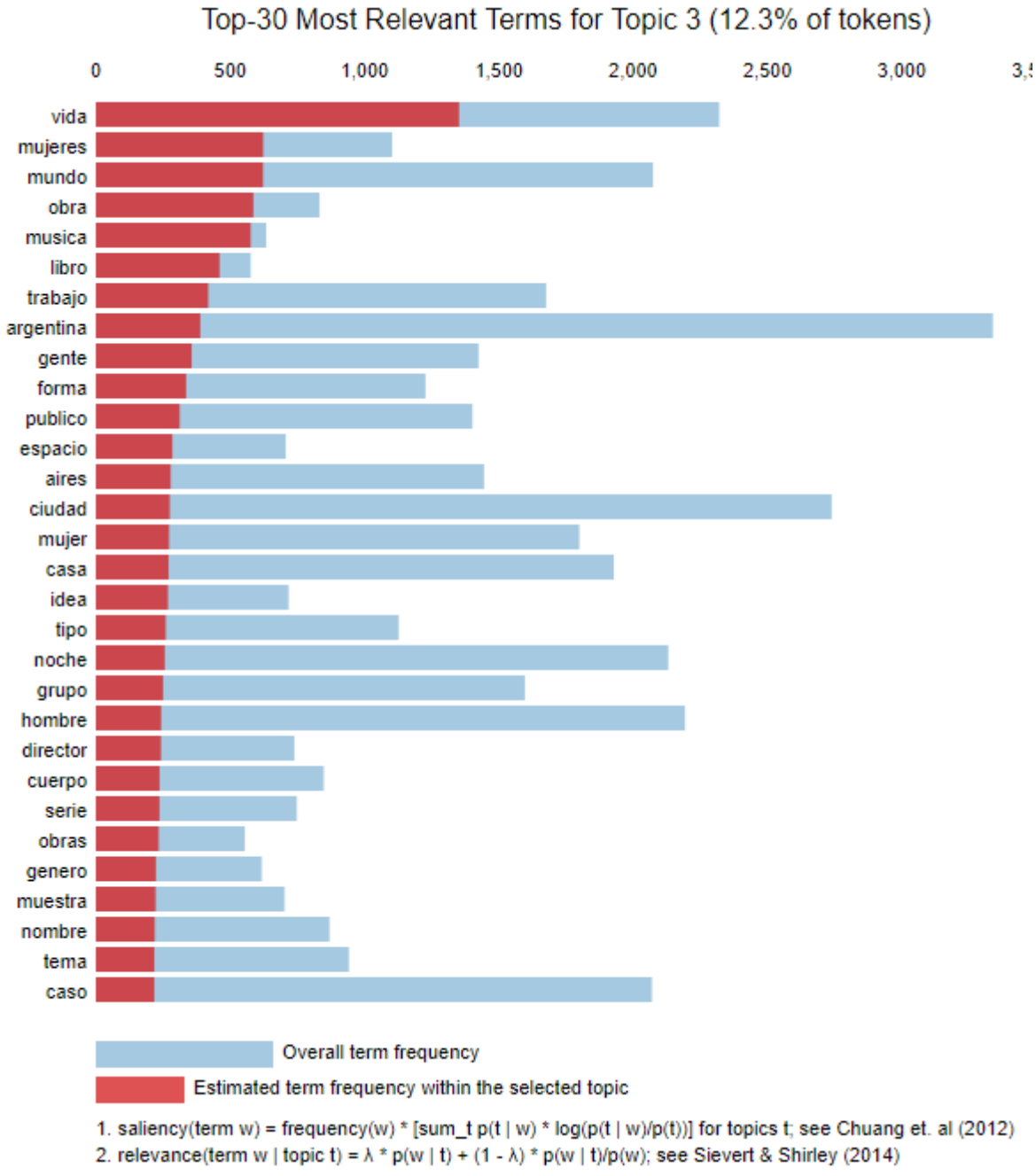
Anexo

I) Tabla de tópicos

Tópico	Nominación
Tópico 1	*Cine, arte y espectáculos*
Tópico 2	*Hechos violentos*
Tópico 3	*Política exterior*
Tópico 4	*Farándula*
Tópico 5	*Causas judiciales y hechos de corrupción*
Tópico 6	*Hechos de inseguridad*
Tópico 7	*Deportes*
Tópico 8	Tópico no interpretable
Tópico 9	*Elecciones*
Tópico 10	*Economía*

II) Tópicos detectados

- *Tópico 1 = "Cine, arte y espectáculos"*



Noticias más representativas del tópico 1:

PERFIL

NOTICIAS

CULTURA | 05-10-2016 22:55

Paulo Coelho: otra vez en los primeros puestos

Acaba de lanzar nueva novela, La Espía, y ya escala posiciones en el ranking de bestsellers. Además, recibió decenas de ofertas por los derechos cinematográficos de su historia.

COMPARTÍ ESTA NOTA →

Secciones LA NACION SUSCRIBITE

LA NACION • Lifestyle

Mesa para dos. Julieta Kemble: "No me gusta que me definan como ex modelo"

Heredera y difusora de la obra de su padre, el artista Kenneth Kemble, acaba de debutar en un reality sobre "las mujeres del polo"

11 de Junio de 2016

Soledad Vallejos

LA NACION

Página|12

Miércoles, 10 de agosto de 2016 | Hoy

ULTIMAS NOTICIAS | EDICION IMPRESA | SUPLEMENTOS | TAPAS | ROSARIO/12

INDICE | EL PAÍS | ECONOMÍA | SOCIEDAD | LA VENTANA | EL MUNDO | ESPECTACULOS | DEPORTES | PSICOLOGÍA | CIENCIA | UNIVERSI

CULTURA & ESPECTACULOS

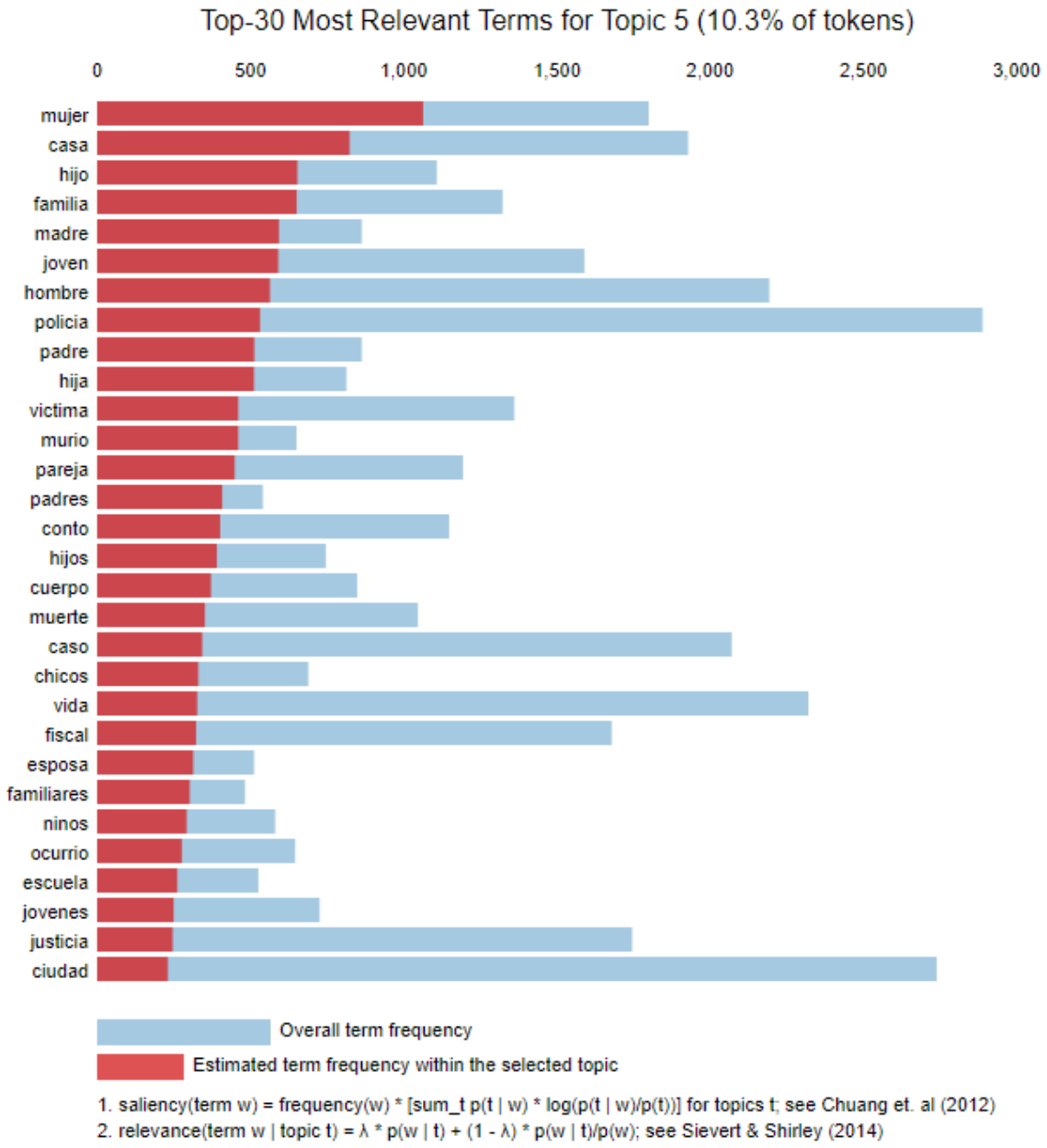
MIÉRCOLES, 10 DE AGOSTO DE 2016

MUSICA › ENRIQUE "ZURDO" ROIZNER SERÁ NOMBRADO PERSONALIDAD DESTACADA DE LA CULTURA

Mucho más que un mercenario del solfeo

Por más que él mismo lo minimice, el aporte del baterista a la música ha sido enorme: fue parte del Octeto Electrónico de Piazzolla, grabó con Vinicius, Gato Barbieri, Saluzzi, Les Luthiers y el Cuarteto Zupay, entre otros. Ahora integra la banda de Kevin Johansen.

- *Tópico 2 = "Hechos violentos"*



infobae

Miércoles 16 de Diciembre de 2020

ARGENTINA

Mató a su esposa y escondió el cuerpo en el baño durante 2 días

Luego, el asesino fue a entregarse a una comisaría de San Juan , acompañado por un sacerdote. Argumentó que lo hizo "en defensa propia". El entorno de la víctima aseguró que es "muy celoso"

27 de febrero de 2016

m1 Secciones Coronavirus Carlos Menem Vacuna

POLICIALES

"Le dieron droga y la abandonaron", dijo el papá de la menor que murió de sobredosis

13 de agosto de 2016

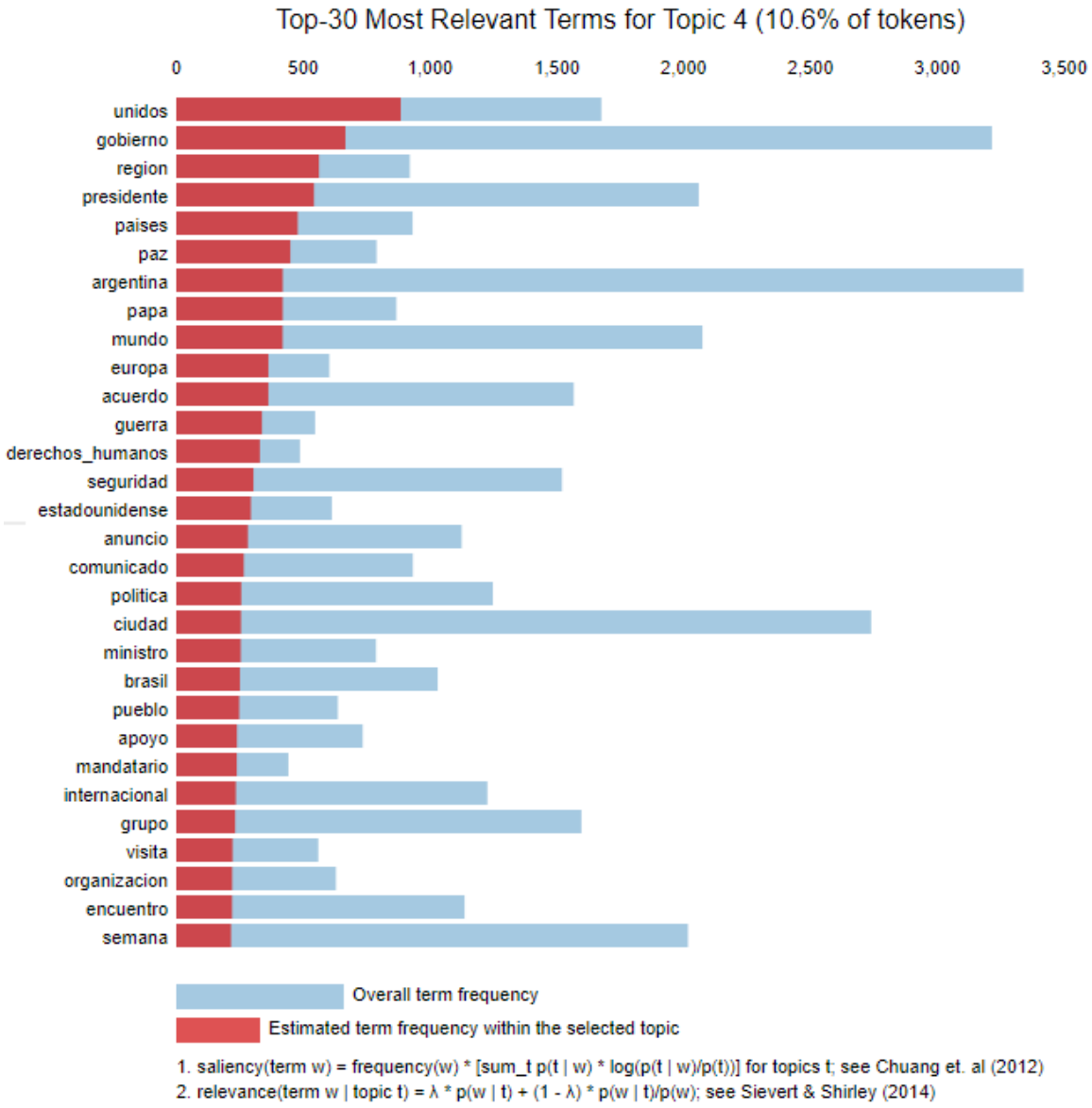
CLARIN.COM ›

Annotations

Las tragedias aéreas provocaron más de 400 muertos este año

CLARÍN.COM DECEMBER 25, 2016

- *Tópico 3 = "Política exterior"*



Noticias más representativas del tópico 3:

télam 75 años

LATINOAMÉRICA

26/09/2016 APOYO

Los generales de la policía colombiana dieron "la bienvenida a la firma de la paz" entre el gobierno y las FARC

Secciones LA NACION SUSCRIBITE

LA NACION • El Mundo

La ONU exige cumplir el alto el fuego en Ucrania tras los últimos combates

En la reunión de emergencia del Consejo de Seguridad, que aborda la crisis, admitieron temer que el conflicto recrudezca en el corto plazo en algunas zonas

5 de Junio de 2015 • 12:06

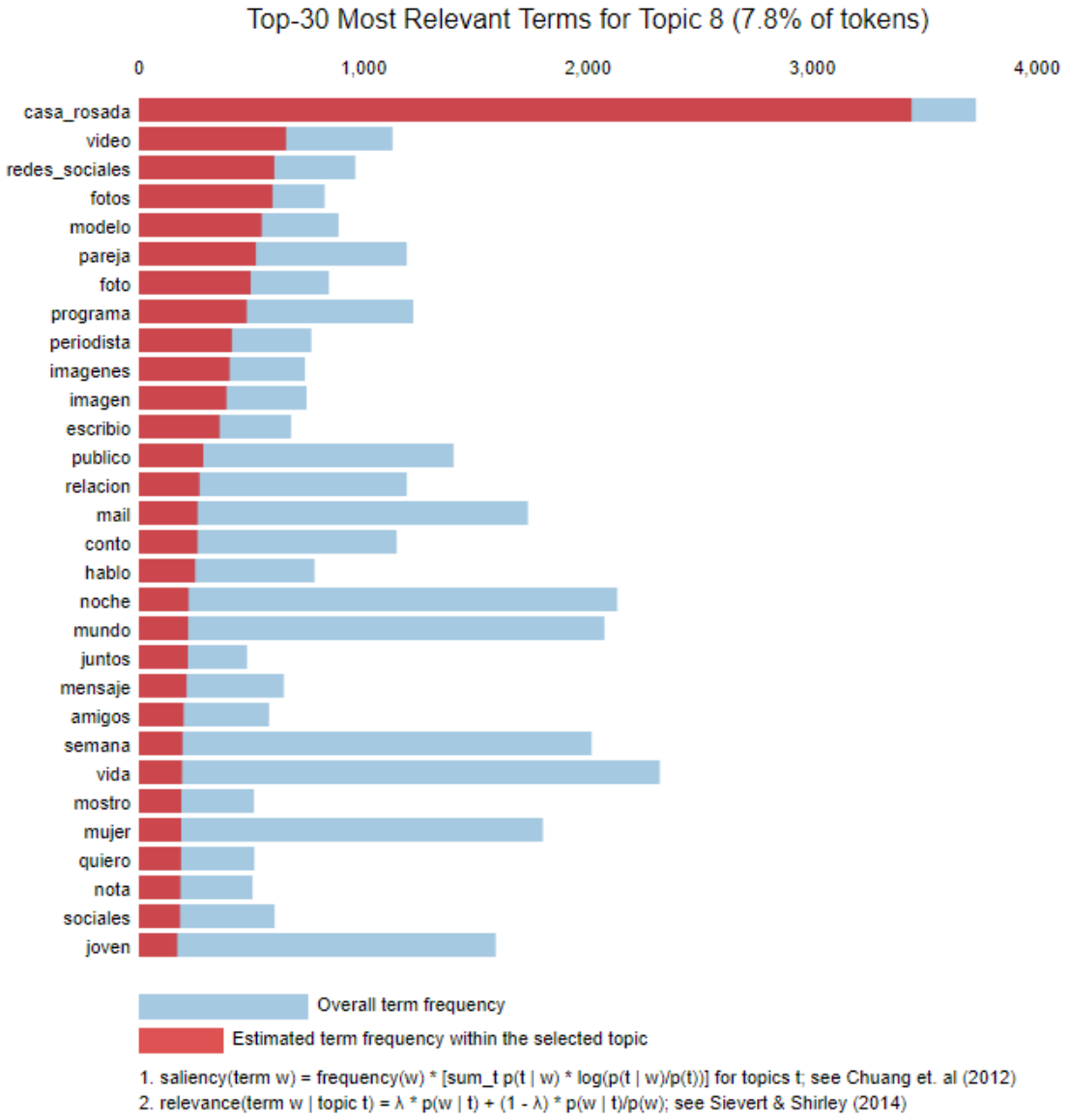
m1 Secciones Coronavirus Carlos Menem Vacuna

PERÚ

La hija de Alberto Fujimori, favorita para llegar a la presidencia de Perú

08 de abril de 2016

- *Tópico 4 = "Farándula"*



Noticias más representativas del tópico 4:

CLARIN.COM ›

Annotations

Gran Hermano 2015: ¿Brian le pegó a Marian?

CLARÍN.COM DECEMBER 08, 2016

m1 Secciones Coronavirus Carlos Menem Vacuna

TWITTER

Brutal insulto de un dirigente del PRO a la periodista Nancy Pazos

25 de noviembre de 2015

infobae

Miércoles 16 de Diciembre de 2020

AMÉRICA TELESHOW DEPORTES TENDENCIAS CULTURA MIX5411

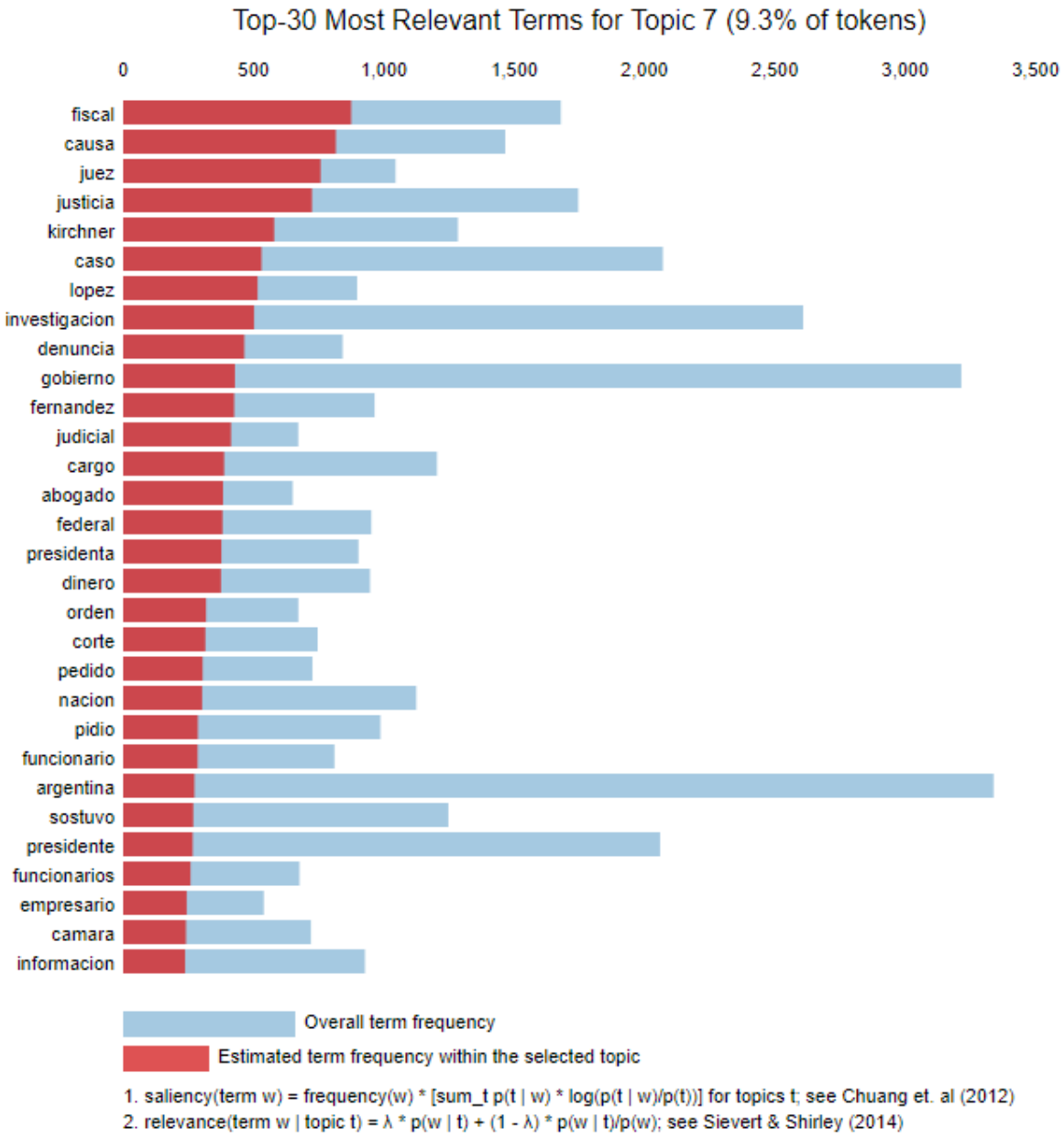
INFOSHOW

La escapada romántica de Fede Bal y Barbie Vélez a Roma

Luego de pasar por una fuerte crisis durante el verano, los actores estuvieron en Italia

21 de marzo de 2016

- *Tópico 5 = "Causas judiciales y hechos de corrupción"*



Noticias más representativas del tópico 5:

PERFIL

INGRESAR SUSCRIBITE

POLITICA | POLMICAS DECLARACIONES

Lunes 21 Marzo, 2016

Miriam Quiroga: "Se dice que Cristina es amante de Alasino"

La exsecretaria personal del expresidente vinculó a la exmandataria con quien fuera jefe del bloque del PJ durante el menemismo.

Secciones

LA NACION

SUSCRIBITE

LA NACION • Política

Aumento de tarifas: habilitan la feria judicial para tratar el tema

18 de Julio de 2016 • 18:59

PERFIL

INGRESAR SUSCRIBITE

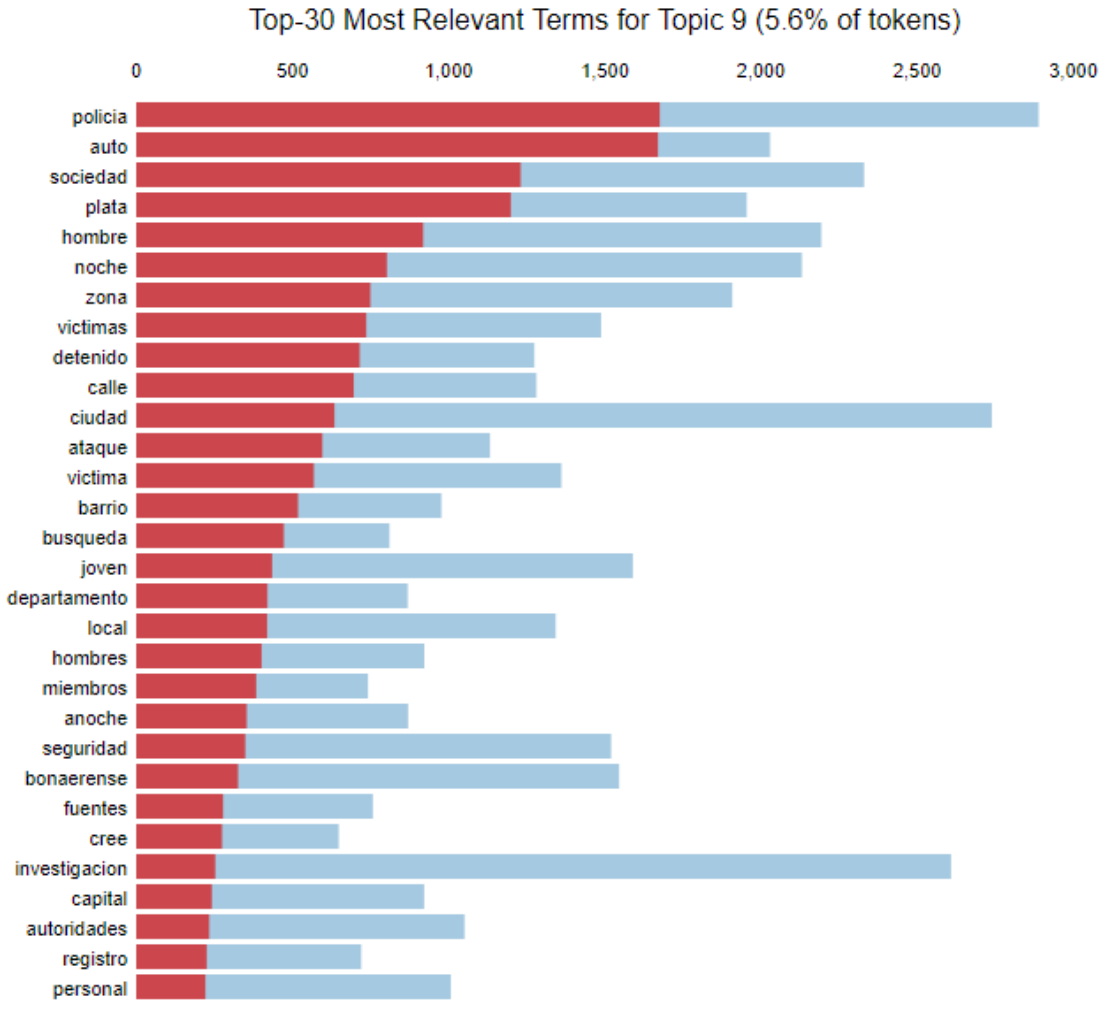
POLITICA | DENUNCIA

Miércoles 21 Octubre, 2015

El hijo de Canicoba Corral, el elegido de Cristina, investigará la causa de espionaje K

Emiliano Ramón Canicoba juró el pasado 30 de junio como titular de un juzgado federal de San Martín. Cómo fue su meteórico ascenso. La relación de su padre con el Gobierno.

- *Tópico 6 = "Hechos de inseguridad"*



Noticias más representativas del tópico 6:

CLARIN.COM ›

Annotations

Lo buscaban por un robo, le encontraron un arsenal

CLARÍN.COM MAY 22, 2016

POLICIALES

Tremendo golpe en la autopista: se llevaron 11 millones de pesos

01 de junio de 2016

Inseguridad

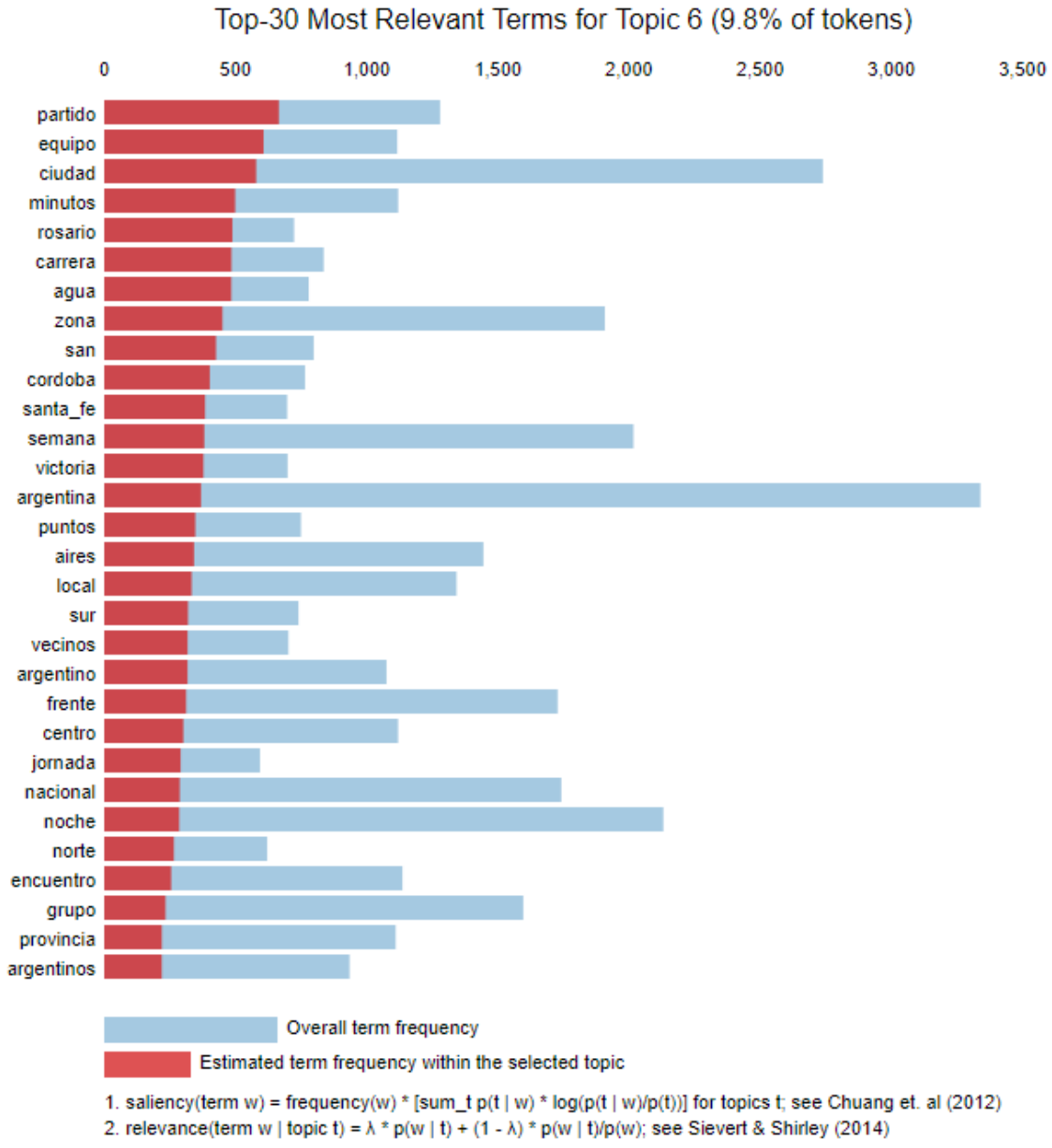
Le pegan tres tiros a un policía para robarle la moto: está muy grave

Ocurrió anoche en San Miguel. El policía iba en su moto vestido de civil cuando lo asaltaron dos ladrones y quiso defenderse.

COMENTARIOS (5)

08/05/2015 9:56 | Clarín.com **Policiales** | Actualizado al 08/12/2016 21:07

- *Tópico 7 = "Deportes"*



Noticias más representativas del tópico 6:

infobae

Miércoles 16 de Diciembre de 2020

AMÉRICA TELESHOW DEPORTES TENDENCIAS CULTURA MIX5411

DEPORTES

El partido de Messi: llegó a su gol 400 en el Barcelona y sigue haciendo historia

Ya había asistido a Luis Suárez en el primero frente al Valencia y más tarde se le había negado en dos ocasiones: una clara en la etapa inicial y un tiro libre en el travesaño en el complemento. Pero con el rival jugado en ataque, aprovechó una contra y, si bien no pudo ponerle un moño, salió favorecido con un rebote y se dio el gusto de celebrar su gol número 400 con la camiseta blaugrana. El hombre récord sigue con su paso arrollador.
REPASÁ EL PARTIDO DE MESSI

Secciones

LA NACION

SUSCRIBITE

LA NACION • Deportes • Fútbol • Huracán

Huracán-Central Córdoba, en vivo: cómo ver online el partido de 32os de final de la Copa Argentina 2016

El encuentro se disputará en el estadio Florencio Sola (Banfield) desde las 21.10, con transmisión por Canal Metro y TyC Sports Play

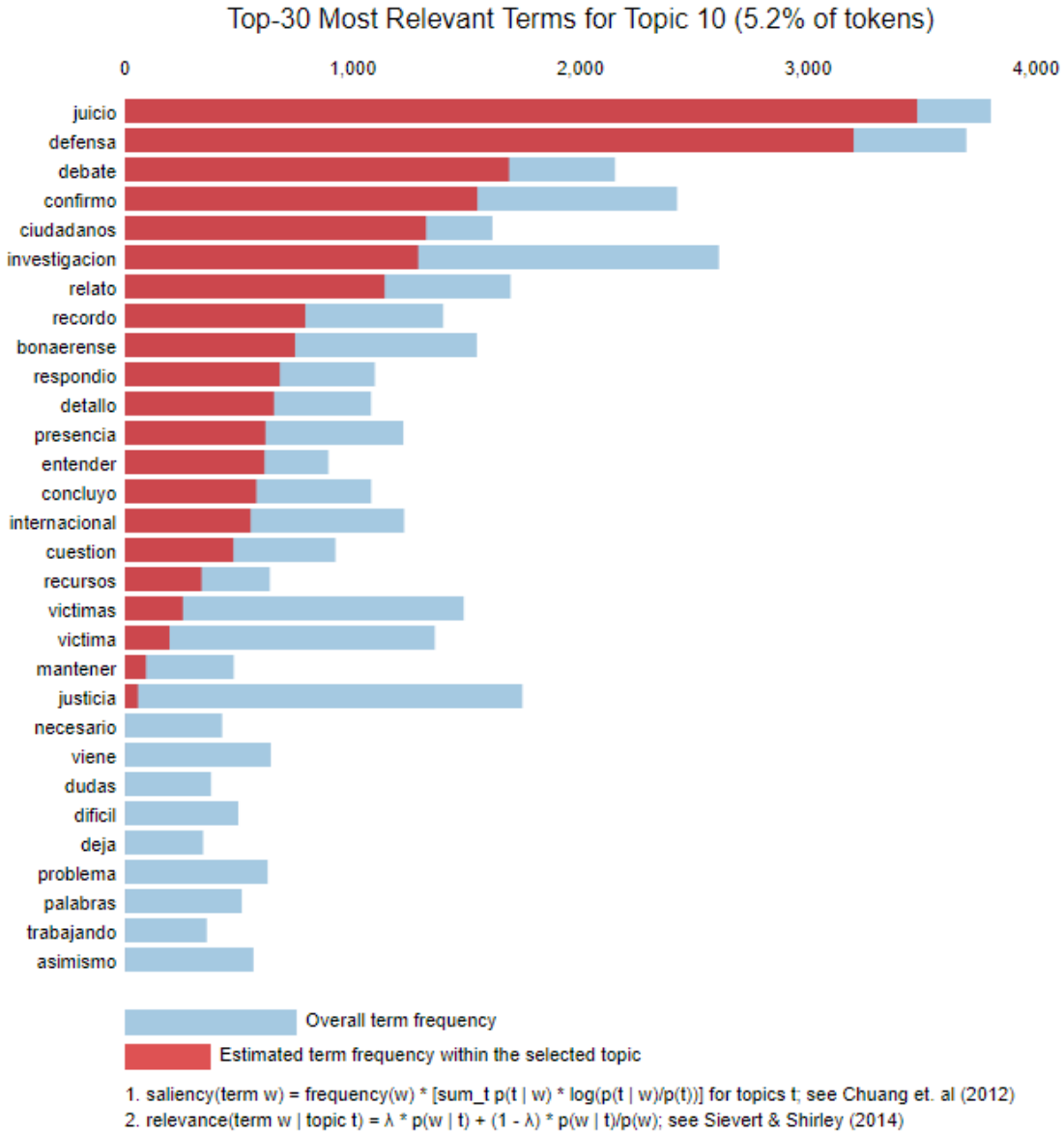
18 de Julio de 2016 • 13:13

BOXEO | 10-08-2016 20:36

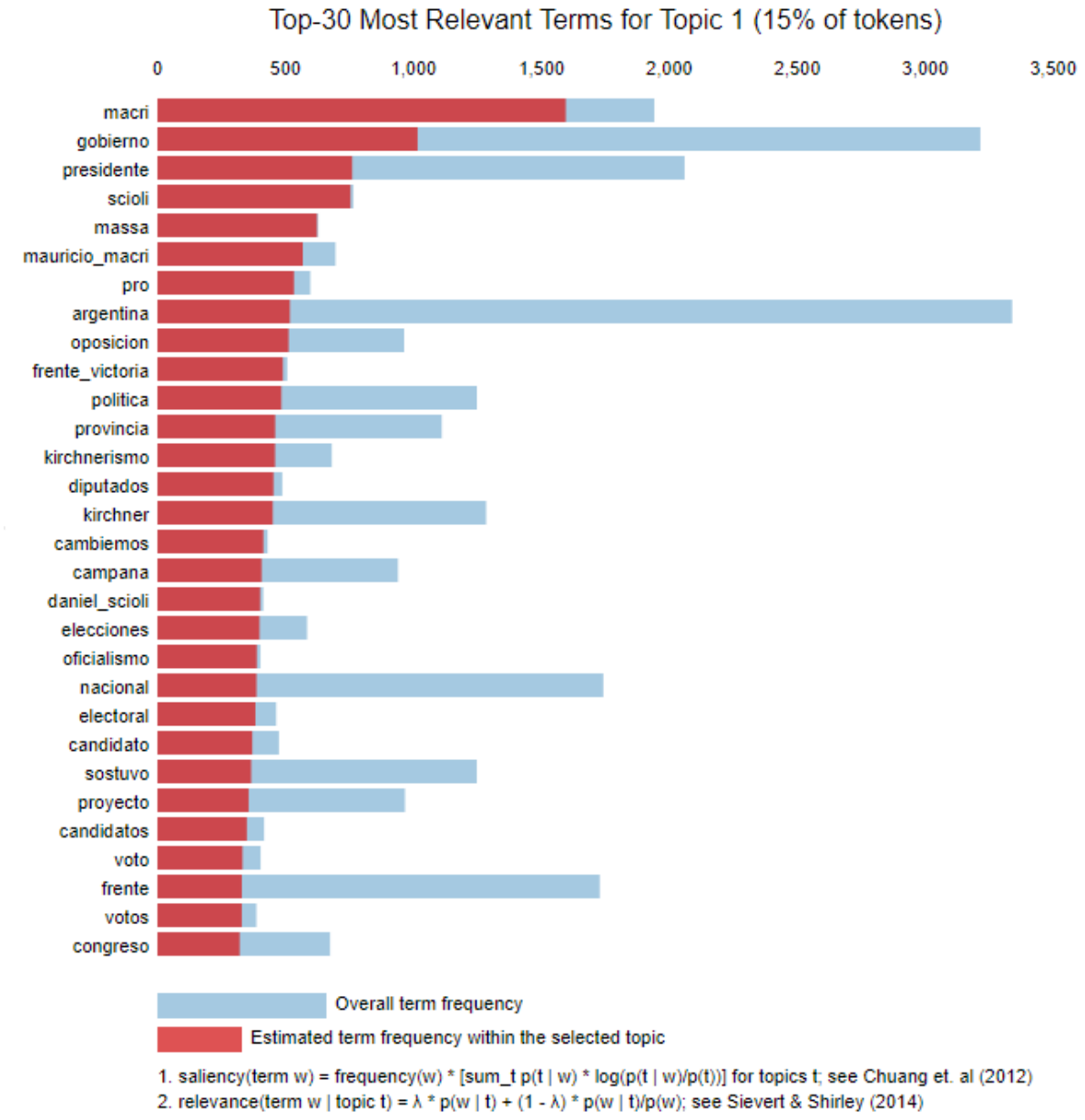
Yamil Peralta no pudo con la categoría de Savón

El boxeador argentino se cruzó con una de las potencias de la categoría hasta 91 kilos, Erislandy Savón, quien lo eliminó y lo dejó sin chances de obtener una medalla.

- *Tópico 8* = Tópico no interpretable



- *Tópico 9 = "Elecciones"*



Noticias más representativas del tópico 9:

Secciones

LA NACION

SUSCRIBITE

LA NACION • Política

Elecciones 2015 / La estrategia oficialista en Santa Cruz. Máximo Kirchner se lanzó como candidato a diputado, pero no habló

El FPV armó un acto con Scioli y Zannini en homenaje a Néstor Kirchner, en el que iban a impulsar la postulación del hijo de la Presidenta; sólo hablaron Alicia Kirchner y los integrantes de la fórmula

3 de Julio de 2015

Pedro Gojan

PARA LA NACION

Secciones

Ingresar Suscribite por $70

ELECCIONES 2015

Lousteau: "El Frente para la Victoria quiso instalar una mentira"

El candidato a jefe de gobierno porteño dijo que la actitud del kirchnerismo de pretender instalar que salieron segundos fue "una muestra de desprecio a los votantes".

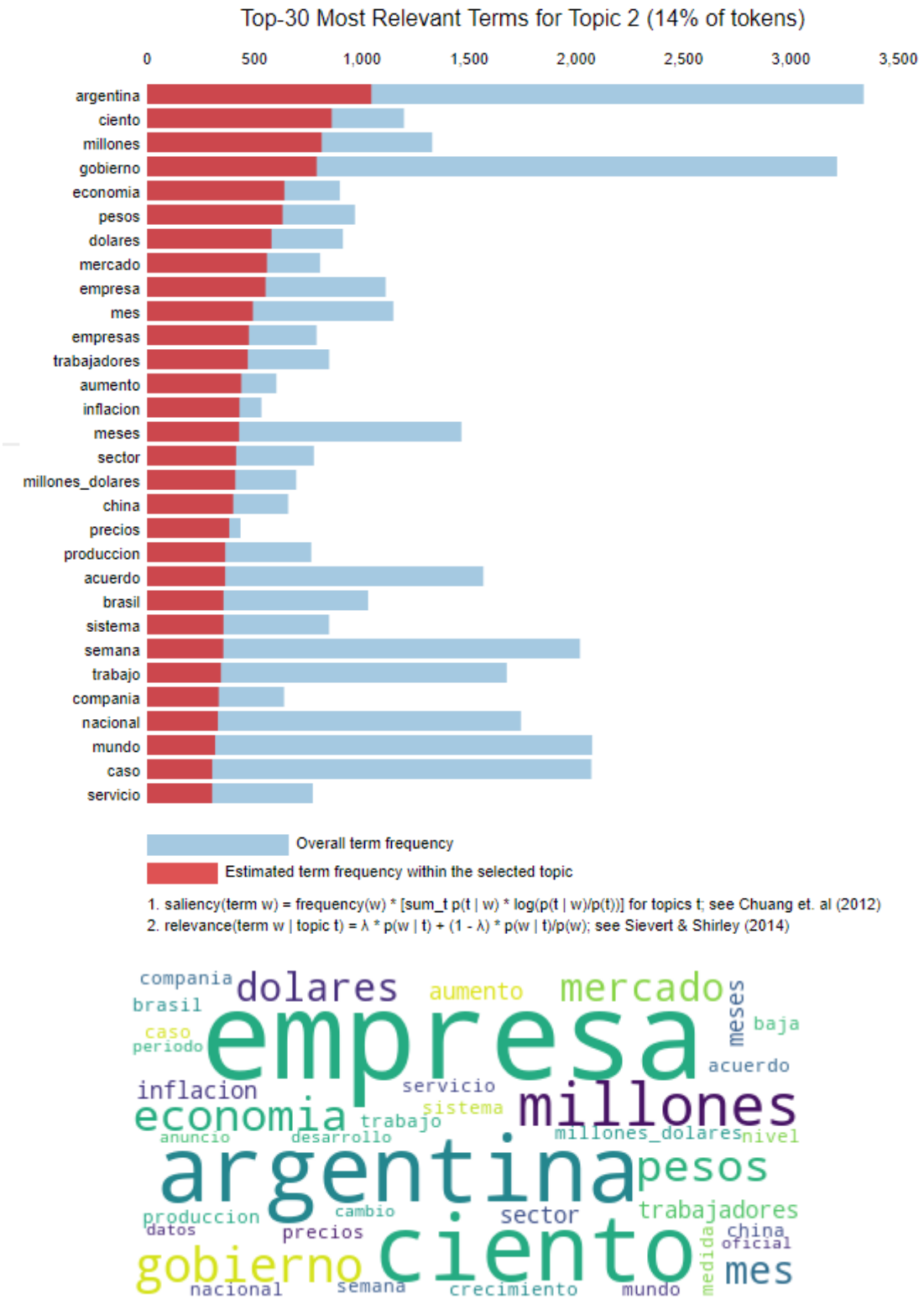
ARGENTINA

Mauricio Macri: "Todos tenemos que poner el hombro para salir de una transición difícil"

El Presidente habló ante un centenar de empresarios y líderes sindicales tras firmar un acuerdo para suspender despidos por 90 días. "Tenemos que debatir en serio para generar trabajo", afirmó

9 de mayo de 2016

- *Tópico 10 = "Economía"*



Noticias más representativas del tópico 10:

infobae

Miércoles 16 de Diciembre de 2020

AMÉRICA TELESHOW DEPORTES TENDENCIAS CULTURA MIX5411

ARGENTINA

Nuevos pagos de bonos amenazan a las reservas

Esta semana se utilizarán divisas del BCRA para cumplir vencimientos de Bonar X, préstamos de organismos internacionales y títulos de la provincia de Buenos Aires. El dólar libre cerró a 15,82 pesos

Por **Luis Beldi** | 14 de octubre de 2015

Secciones LA NACION SUSCRIBITE

LA NACION • Política

El FMI apuesta a que la Argentina aporte estabilidad a América latina

Elogió las medidas del Gobierno, aunque admitió que el ajuste "será difícil de digerir"

17 de Abril de 2016

Parabrisas 60 AÑOS

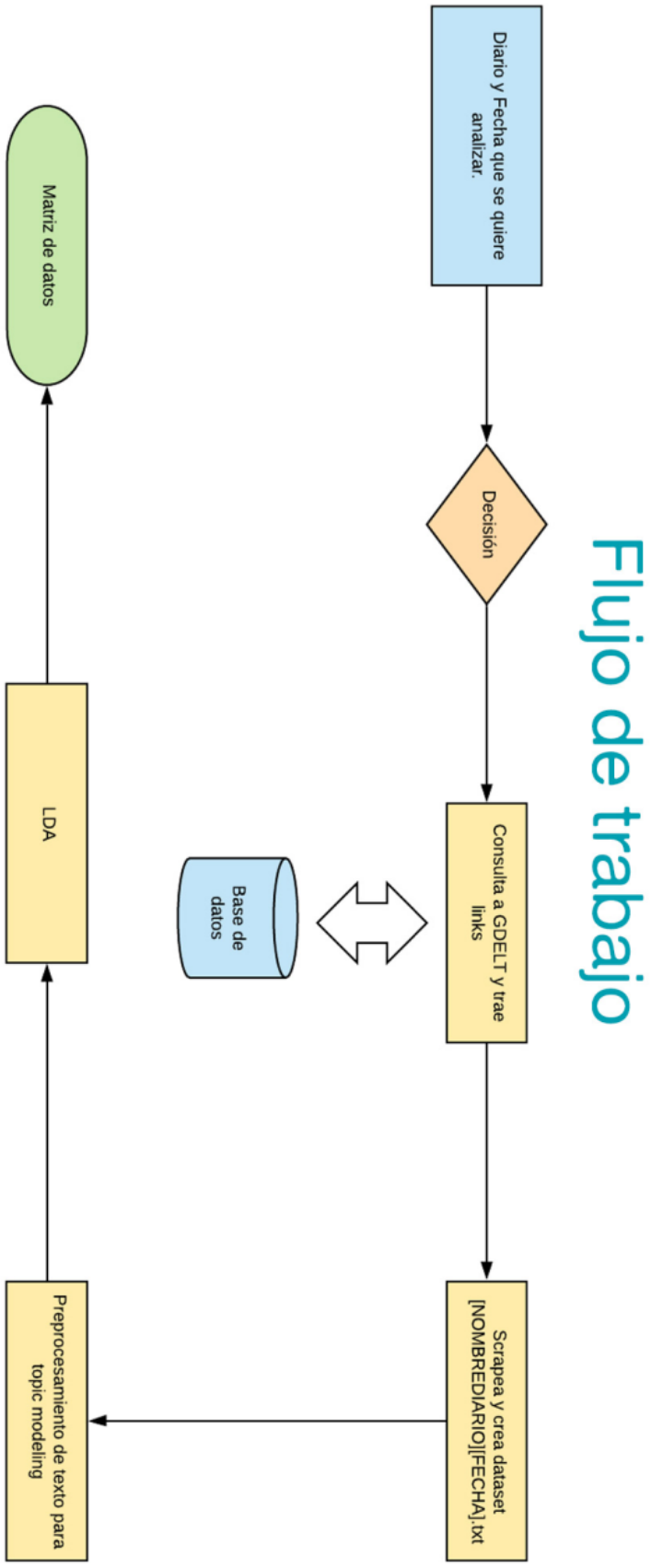
ÚLTIMO MOMENTO

MERCADO | 06-04-2016 12:42

Fiat anunció inversiones por u$s500 millones

La filial nacional de la automotriz destinará ese monto a su planta de Ferreyra, Córdoba, para el desarrollo de un nuevo modelo para la exportación regional.

III) Diagrama de flujo



Bibliografía citada

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. Wired, 23 June 2008. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

Anyoha, R. (2017) "The History of Artificial Intelligence", Harvard Universiy Press. url: http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Aruguete, N. (2015) "El poder de la agenda: Política, Medios y Público", Cuadernos de comunicación, Ed. Biblos.

Becerra, M. (2019): Medios digitales en Argentina: la película y la foto. Recuperado el 20 de septiembre de 2019, de: https://www.letrap.com.ar/nota/2018-9-20-16-3-0-medios-digitales-en-argentina-la-pelicula-y-la-foto

Bird, Klein y Loper (2009) "Natural language processing with Python: Analyzing text with the natural language toolkit", O'Reilly, California, EE.UU.

Breiman, L. (2001) "Statistical Modeling: The Two Cultures", Statistical Science Vol. 16, No. 3, 199-231. Berkeley, California, EEUU.

Calvo, E. (2015) "Anatomía política de Twitter. Tuiteando a #Nisman". Capital Intelectual. Bsas. Argentina.

Calvo E. y Aruguete, N. (2018) "#Tarifazos. Medios tradicionales y fusión de agenda en redes sociales". Inmediaciones de la comunicación 2018 – VOL. 13/N° 1.

Carbonell J, Michalski R, Mitchell T. (1983) "Machine Learning: A Historical and Methodological Analysis", AI Magazine Volume 4 Number 3 1983, Palto Alto CA.

Chollet, F. (2018) "Deep Learning with Python" Manning Publications, Shelter Island, NY.

Cioffi – Revilla, C. (2010) "Computational social science", WILEY Interdisciplinary Reviews: Computational Statistics, Vol. 2, no. 3, Mayo/Junio 2010:pp. 259–271.

DiMaggio P, Nag M y Bleid, D (2013), Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics 41(6): 570 – 606.

Eilders, C. (1997), “The impact of editorial content on the political agenda in Germany: Theoretical assumptions and open questions regarding and neglected subject in mass communication research”, Discussion Paper FS III, Berlin.

- (2000), “Media as political actors? Issue focusing and selective emphasis in the German Quality Press”, German Politics, 9 (3): 181 – 206.

Gálvez, R.H., Tiffenberg, V. & Altszyler, E. (2019) “Half a Century of Stereotyping Associations Between Gender and Intellectual Ability in Films”, Sex Roles 81, 643–654. https://doi.org/10.1007/s11199-019-01019-x

Garg N., Schiebinger L., Jurafsky D., Zou J. (2018) “Word embeddings quantify 100 years of gender and ethnic stereotypes”, Princeton University, Princeton, NJ, EEUU.

Hastie, R. y Tibshirani, R. (1990) Generalized Additive Models, Chapman & Hall/SRC, Nueva York, EEUU.

Hilbert, M., & López, P. (2011). The World’s Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60 –65. doi:10.1126/science.1200970

Kitchin, R. (2014) “Big data, new epistemologies and paradigm shifts”, Big Data & Society, Abril – Junio 2014: 1 – 12.

Koziner, N (2020) “Temas y fuentes en medios argentinos. Un estudio en contexto electoral (2019)” Más Poder Local. ISSN: 2172-0223. Número 40, Enero 2020, pp.46-56.

Kushin, M.J. (2010), “Tweeting the issues in the Age of Social Media? Intermedia agenda setting between The New York Times and Twitter”, Washington State University.

Lettier, D. (2018) “Your guide to Latent Dirchler Allocation”, Medium. https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7

López-López, P y Vásquez-González, J. (2018) “Agenda temática y Twitter: Elecciones presidenciales en América Latina durante el período 2015-2017”. El profesional de la información, noviembre-diciembre, v.27, n°6. elSSN: 1669-2407.

McFarland, Goldberg y Lewis (2015) “Sociology in the Era of Big Data: The Ascent of Forensic Social Science”, Springer Science + Business Media, New York.

Meraz, S. (2009), “Is there an elite hold? Traditional media to social media agenda setting influence in blog networks”, Journal of Computer-Mediated Communications, 14 (3): 682-707.

- (2011), “Using time series analysis to measure intermedia agenda setting influence in traditional media and political blog networks”, Journalism & Mass Communication Quarterly, 88 (1): 2011.

Mutzel, S. (2015) “Facing Big Data: Making sociology relevant”, Big Data & Society, July – December 2015: 1 – 4. Sage.

Pan, Z. y Kosicki, G. (1993): "Framing analysis: An approach to news discourse". Political Communication, 10(1): 55-75. https://doi.org/10.1080/10584609.1993.9962963.

Rodríguez Zoya y Roggero, «La modelización y simulación computacional como metodología de investigación social», Polis [En línea], 39 | 2014, Publicado el 23 enero 2015. URL: http://journals.openedition.org/polis/10568

Rosati, G. (2017) “Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de Ensamble Learning. Aplicación a la Encuesta Permanente de Hogares (EPH)”, SaberES. Revista de Ciencias Económicas y Estadística 9(1). http://dx.doi.org/10.35305/s.v9i1.132

- (2019). Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data. Simposio llevado a cabo en el workshop de Factor DATA, Universidad Nacional de San Martín, Instituto de Altos Estudios Sociales.

- (2020). “Procesamiento de Lenguaje Natural aplicado a las ciencias sociales. Detección de tópicos en letras de tango”, *Relmis*, en prensa.

Sarraute, C.; Blanc, P. y Burroni, J. (2014). “A study of age and gender seen through mobile phone usage patterns in Mexico”, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 836-843. 10.1109/ASONAM.2014.6921683

Silge, J. y Robinson, D. (2017) “Text mining with R”, O’Reilly, EEUU.

Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S. (2012) “A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle”, In Proceedings of the ACL 2012 System Demonstrations (ACL '12). Association for Computational Linguistics, USA, 115–120.

Weaver, W. (1948) “Science and complexity”, American Scientist. 1948; 36:536-544. https://www.jstor.org/stable/27826254

Zunino, E. y Focas, B. (2018) “El tratamiento informativo de la ‘inseguridad’ en Argentina: víctimas, victimarios y demandas punitivas”. Communication & Society, 2018. Vol 31 (3) pp. 189-209.

Zunino E y Grilli Fox A (2019), “Medios digitales en la Argentina: posibilidades y límites en tensión”, Estudios sobre el Mensaje Periodístico 26(1), 401-413.

Zunino, E. y Ortíz Marín, M. (2016) “Los medios y las elecciones: la agenda informativa de la campaña presidencial de 2015 en la Argentina”. Más Poder Local. ISSN: 2172-0223. Número 30, enero 2017, pp. 56-66.

Bibliografía de referencia

Kozlowski, D. (2019). Análisis computacional de texto aplicado a las ciencias sociales. Aprendizaje automático y Big Data. Simposio llevado a cabo en el workshop de Factor DATA, Universidad Nacional de San Martín, Instituto de Altos Estudios Sociales. https://diegokoz.github.io/workshop_text_mining/1_explicacion.nb.html